ACTA UNIVERSITATIS UPSALIENSIS
Studia Linguistica Upsaliensia
1

Recycling Translations

Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing

BY

JÖRG TIEDEMANN

ACTA UNIVERSITATIS UPSALIENSIS
UPPSALA 2003

Dissertation at Uppsala University to be publicly examined in the lecture hall IX, University Hall, Friday, December 12, 2003 at 10:15 for the Degree of Doctor of Philosophy. The examination will be conducted in English

Abstract

Tiedemann, J. 2003. Recycling Translations. Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. Acta Universitatis Upsaliensis. *Studia Linguistica Upsaliensia* 1. 130 pp. Uppsala. ISBN 91-554-5815-7

The focus of this thesis is on re-using translations in natural language processing. It involves the collection of documents and their translations in an appropriate format, the automatic extraction of translation data, and the application of the extracted data to different tasks in natural language processing.

Five parallel corpora containing more than 35 million words in 60 languages have been collected within co-operative projects. All corpora are sentence aligned and parts of them have been analyzed automatically and annotated with linguistic markup.

Lexical data are extracted from the corpora by means of word alignment. Two automatic word alignment systems have been developed, the Uppsala Word Aligner (UWA) and the Clue Aligner. UWA implements an iterative "knowledge-poor" word alignment approach using association measures and alignment heuristics. The Clue Aligner provides an innovative framework for the combination of statistical and linguistic resources in aligning single words and multi-word units. Both aligners have been applied to several corpora. Detailed evaluations of the alignment results have been carried out for three of them using fine-grained evaluation techniques.

A corpus processing toolbox, Uplug, has been developed. It includes the implementation of UWA and is freely available for research purposes. A new version, Uplug II, includes the Clue Aligner. It can be used via an experimental web interface (UplugWeb).

Lexical data extracted by the word aligners have been applied to different tasks in computational lexicography and machine translation. The use of word alignment in monolingual lexicography has been investigated in two studies. In a third study, the feasibility of using the extracted data in interactive machine translation has been demonstrated. Finally, extracted lexical data have been used for enhancing the lexical components of two machine translation systems.

Keywords: word alignment, parallel corpora, translation corpora, computational lexicography, machine translation, computational linguistics

Jörg Tiedemann, Department of Linguistics. Uppsala University. Villavägen 4, Box 527, 751 20 Uppsala

ISBN 91-554-5815-7
ISSN 1652-1366
urn:nbn:se:uu:diva-3791 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-3791)

Acknowledgments

This work would not have been possible without the help and advice by a large number of colleagues and friends. It is hard to write appropriate acknowledgments that reflect the assistance and inspiration I was given throughout the years of my research.

First of all, I would like to thank my adviser, Anna Sågvall Hein, for her continuous support of my work and for all the hours spent on the improvement of this thesis. Much of the research that is presented here would not have been possible without the encouragement from my supervisor.

Many more people have contributed to this thesis in one way or another. I am grateful for the support by former and present members of the Department of Linguistics in Uppsala. I would like to thank Erik Tjong Kim Sang especially for starting my interest in parallel corpora. Without his supervision during my diploma thesis many years ago I would not be here at this point presenting continued research in the field. Thanks to Lars Borin for fruitful discussions and scientific insights. I miss the meetings of the corpus interest group since he left Uppsala. Many thanks also to Leif-Jöran Olsson for running my scripts and giving me valuable feedback. I am also grateful for the administrative and technical support at the department that made it possible to carry out the work I am presenting. Per Starbäck knows how to handle our computer system and Gunilla Fredriksson knows how to handle forms and applications. I am also indebted to Per Weijnitz for showing me the beauty of remote computing. It has been a great pleasure to know that 20 computers work over night and produce results for my evaluation. Thanks to various people for comments and suggestions. Mats Dahllöf told me how to skip the boring parts in the drafts of my thesis and to come right to the point in my writings. Bengt Dahlqvist checked my formulas and gave me a large number of photo copies about his favorite measures in statistics. Thanks to other people who read parts of my thesis, who reviewed the papers the thesis is partly based on, and who gave me the possibilities to present my ideas.

I would also like to thank the people who worked together with me in various projects during the past years. Thanks especially to Lars Ahrenberg, Magnus Merkel, and Mikael Andersson for all inspiring discussions and contributions to my work . Thanks also to Eva Forsbom, Ebba Gustavii, Eva Pettersson, Michael Petterstedt, Maria Holmqvist, Ingrid Almqvist, Daniel

Ridings, Katarina Mühlenbock for the fruitful co-operations. I am especially grateful for the fantastic ideas of Lars Nygaard. I am delighted to be a member of the OPUS project and hope for a long continuation of this co-operation.

I am indebted to the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) for a generous grant that made it possible to experience one of the finest research institutes in the world at the University of Edinburgh. People I have met there taught me a lot that contributed directly or indirectly to this thesis.

Many of my friends should be mentioned here. The biggest acknowledgments go to Emma who had the energy to work through my text and to polish my English. Thank you very much. I cannot list all the people who made my life in Uppsala worthwhile. Doubtless, without friends I would never have managed to get through this work.

Finally, and most importantly, my deepest thanks to the most wonderful person in my life, Therese, for all her patience with me and my changing moods. To Therese, with all my love!

Jörg Tiedemann
Uppsala, October 2003

Contents

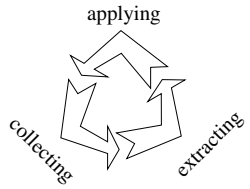
List of Tables

List of Figures

1 Introduction

The title of this thesis sets the focus on re-using translations in natural language processing, a task which can be seen in a recycling framework. The idea of using previous translations for new tasks is not new. Translated texts have been used as reference data in human translation and language teaching for many years. However, the significance of such data has grown enormously with the development of computer technology that revolutionized the way of processing natural languages. Computational linguists discovered the potentials of previous translations in computational lexicography and machine translation a couple of decades ago. Since then, much work has been devoted to building representative collections of documents and their translations, and to the development of tools for processing such collections. The aim of the thesis is to investigate techniques in this field of research, to develop them further, and to explore possible applications.

In general, recycling involves three tasks: the collection of material, the extraction of re-usable components, and the refinement and application of extracted items.



This thesis deals with all three aspects in recycling translations: the compilation of *parallel corpora*[1] from previous translations, the extraction of bilingual lexical data from these corpora, and the application of extracted data to tasks in natural language processing. Contributions of the thesis include a comprehensive collection of parallel corpora, innovative techniques for the automatic alignment of translation data in such corpora and their extraction, fine-grained methods for the automatic evaluation of lexical alignment, and a set of tools for processing parallel corpora and handling multilingual lexical data. The thesis also includes a detailed evaluation of the alignment techniques that have been proposed, and finally, four examples of applications of extracted lexical data in computational lexicography and machine translation.

[1] The concept of parallel corpora will be explained in section 2.1.

1.1 Aims and objectives

The primary goal of this thesis is to develop and apply computational techniques for the extraction of translation data from previous translations. The focus is set on written data involving Swedish either as source or target language. However, with minor adjustments, these techniques should also be applicable to other language pairs.

In pursuing the main goal, the following tasks have been distinguished:

Data collection: Documents and their translations have to be collected and transformed into a format which can be used extensively and efficiently. This task involves format conversions and basic pre-processing such as sentence splitting, tokenization, and linguistic tagging.

Development of tools: Data without tools cannot be processed nor "recycled". Tools in terms of computer programs have to be implemented for working with the data collection. This task includes the development of systems for general corpus processing, for lexical extraction, and for handling extracted data.

Alignment and extraction of lexical data: Alignment is the fundamental task in the research presented. The most important feature of texts and their translations is the correspondence between source and target segments. By alignment we understand a process of linking such corresponding segments in translation data. This can be carried out at several segmentation levels, such as paragraphs, sentences and words (see sections 2.2 and 2.3). The focus here is on the alignment at the lexical level, i.e. the linking of words and phrases. Automatic alignment is often incomplete and never perfect. Hence, part of this task is the evaluation of alignment results. Evaluation methodologies have to be investigated and techniques for the systematic evaluation of alignment results have to be developed.

Applications: The final task comprises the application of extracted data to tasks in natural language processing. Two fields of research are examined: computational lexicography and machine translation.

1.2 Outline

The thesis includes seven chapters presenting research that has been carried out in recent years. Parts of the thesis elaborate work by the author that has been published elsewhere; references will be given in the text. Other parts, especially chapter 5, contain recent, unpublished work that is described in detail in comparison with earlier achievements.

Chapter 2 (which follows this introduction) provides some background to the field of research on translation corpus processing. It introduces basic terminology and includes a summary of related work. It presents common sentence alignment techniques, word alignment models, and a discussion of evaluation techniques for the latter. Furthermore, some related projects that apply the presented approaches are listed in the end of this chapter.

Chapter 3 gives an overview of translation corpora, which have been collected, built and used in the thesis. It contains a brief introduction to corpus encoding and a summary of characteristics for each of the corpora.

Chapter 4 describes tools that have been implemented in the thesis. It includes a description of corpus processing tools and an overview of database tools for storing lexical data.

Chapter 5 constitutes the main contribution of the thesis. Here, our automatic word alignment techniques are presented and discussed. Two systems are described, an iterative "knowledge-poor" approach based on association measures and alignment heuristics (the Uppsala Word Aligner), and a probabilistic alignment framework for the combination of statistical and linguistic resources (the Clue Aligner). The chapter also includes a proposal of refined evaluation measures for word alignment, and a detailed presentation of recent alignment experiments using the techniques described in the thesis.

Chapter 6 includes an overview of four studies on the application of alignment results to tasks in the field of computational lexicography and machine translation.

Chapter 7 concludes the thesis with a summary of contributions and some prospects for future work.

1.3 Project framework

The research presented in this thesis is mainly based on work that has been carried out in several research projects on parallel corpora and machine translation. Goals, run-time, funding and organization differ for each of the projects. However, common to all of them is the use of parallel corpora as their main source of data.

The following list includes brief descriptions of the projects involved:

PLUG: *Parallel Corpora in Linköping, Uppsala and Göteborg* (PLUG), a co-operative project aimed at the development, evaluation and application of programs for alignment and data generation from parallel corpora with Swedish as either source or target language. The participating departments were the Department of Swedish at Göteborg University, the Department of Computer and Information science at Linköping University, and the Department of Linguistics at Uppsala University. The project was funded by NUTEK (The Swedish Board of Industrial and Technical Development) and HSFR (The Swedish Council for Research in the Humanities). PLUG began in autumn 1997 and was finished in the end of 1999. More about aims and achievements of this project can be found in [Såg02].
http://stp.ling.uu.se/plug/

MATS: *Methodology and Application of a Translation System* (MATS), a joint project carried out in co-operation between the Department of Linguistics at Uppsala University, Scania CV AB, and Translator Teknikinformation AB. This project was co-ordinated by Uppsala University and funded by NUTEK (The Swedish Board of Industrial and Technical Development) as part of the FavorIT program. The project was started in October 2000 and ended in May 2001.
http://stp.ling.uu.se/mats/

KOMA: *Corpus-based machine translation* (KOMA) is a project within the VINNOVA[2] research program for language technology. The aim of the project is to develop methods and systems for machine translation of documents of a restricted text type. The project is an on-going joint project between the Natural Language Processing Laboratory (NLPLAB) at the Department of Computer and Information Science, Linköping University, and the Department of Linguistics at Uppsala University.
http://www.ida.liu.se/ nlplab/koma/

[2]Swedish Agency for Innovation Systems

Contrastive Lexicology and Recognition of Translation Equivalents: This project is carried out within the joint research program *Translation and Interpreting. A Meeting between Language and Culture* (Stockholm University and Uppsala University). It is financed by The Bank of Sweden Tercentenary Foundation (Riksbankens jubileumsfond).
http://www.translation.su.se/

Scania: The Scania project is a long-term co-operation between Scania AB in Södertälje and the Department of Linguistics at Uppsala University. The project aims at terminology extraction, language control and machine translation of technical manuals.
http://stp.ling.uu.se/scania/

OPUS: The *Open Source Parallel Corpus* is an initiative by Lars Nygaard from the Text Laboratory at the Faculty of Arts at the University of Oslo and Jörg Tiedemann from the Department of Linguistics at Uppsala University. The project was started in December 2002 and does not have any funding.
http://logos.uio.no/opus/

This thesis summarizes the overall research results without relating each of them to specific projects. Hence, it does not present all its parts in a chronological order. Furthermore, two projects (KOMA and OPUS) are on-going and their results will exceed the ones described here.

2 Background

In this chapter, I will introduce basic concepts and techniques in the field of translation corpus processing. The chapter includes definitions of basic terminology, as they are used in the thesis, and a summary of standard methods for the compilation of *parallel corpora* and their application to computational lexicography and machine translation. Figure 2.1 gives an overview of the compilation and use of parallel corpora within the field of natural language processing.

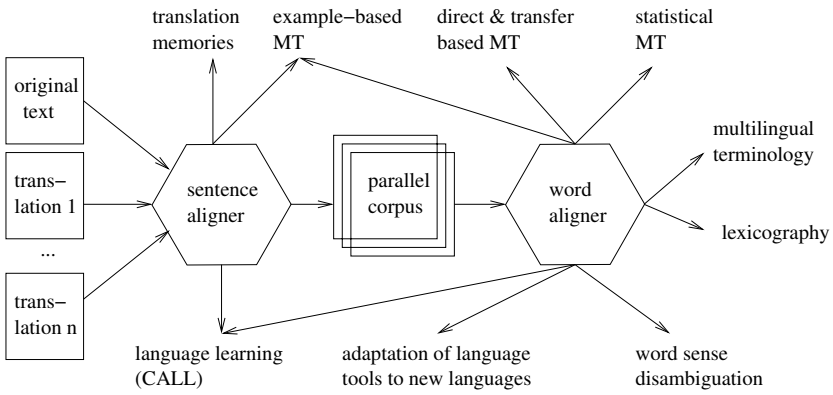


Figure 2.1: Compiling and applying parallel corpora.

The following section (2.1) defines the term *parallel corpus* as used in the thesis in relation to other concepts of computational corpus linguistics. This is followed by a short overview of common sentence alignment techniques in section 2.2. The main part of this chapter is devoted to *word alignment* (section 2.3). In this part, I will briefly introduce the concept of word alignment, review two general approaches thereof and finally, discuss evaluation methodologies of alignment results. Some applications of these techniques are listed in section 2.4. Finally, the last part of this chapter links background information to other parts of the thesis.

2.1 Parallel corpora

In computational linguistics, a corpus is a (in some respect representative) collection of spoken or written utterances of natural language usually accessible in electronic form. Often, corpora represent a particular genre of text or speech. Other corpora contain a large variety of types and genres to represent language use in a more general way. However, a corpus is always just a sample and can never completely represent a whole language. The expressive power of natural language cannot be captured by a finite data set.

There are several ways of classifying corpora into different types and categories according to their properties. One way is to distinguish between corpora that include only one language (monolingual corpora) and corpora that include several languages (multilingual corpora). Multilingual corpora can be divided into parallel and non-parallel corpora. Parallel corpora are referred to as natural language utterances and their translations with alignments between corresponding segments in different languages[1]. The alignment distinguishes parallel corpora from other multilingual corpora such as general translation corpora or so-called *comparable corpora*. Parallel corpora usually contain a common source document (the original) and one or more translations of this source (target documents). Sometimes the original language is unknown (*mixed source corpora*) or the original document is not included at all (*multi-target corpora*) [Mer99b]. Bilingual parallel corpora are sometimes called *bitexts* (see e.g., [Isa92]) and corresponding parts within these corpora are called *bitext segments* (see e.g., [AMST99]).

Parallel corpora have been exploited in many studies. Many applications use parallel corpora for translation studies and for tasks in multilingual natural language processing (NLP). Bilingual concordances have been used for some years in order to support human translation. In recent years, parallel corpora have become more widely available and serve as a source for data-driven NLP tasks. Automatic extraction of multilingual term databases, statistical machine translation, corpus-based bilingual lexicography are just some research fields that have been developed in connection with a growing number of large parallel corpora.

The most widely used parallel corpora are derived from the English and French records of the Canadian Parliament, the so-called *Hansards* corpora. A compilation of these records are available from, for instance, *http://www.isi.edu/natural-language/download/hansard/index.html*. Like the Hansards, most parallel corpora contain only two languages, a source and a target language. However, multilingual parallel corpora with translations

[1] Here, we refer to parallel corpora exclusively in terms of multilingual parallel corpora. Other types of parallel corpora include diachronic corpora (different versions of the same document from different periods of time) and transcription corpora (e.g. textual representations of spoken language or dialects aligned to a corresponding standard language text). [Mer99b]

into more than one language are available and became very popular in recent studies. Examples of such corpora are the Multext East "1984" corpus[2] for central and eastern European languages, the multilingual parallel corpus of European Parliament proceedings EUROPARL[3] in eleven languages and the multilingual OPUS corpus[4], which is briefly described in section 3.2.5.

2.2 Sentence alignment

Source language documents in a translation corpus can be split into segments that correspond monotonically to segments in translated documents. Common segmentation units are paragraphs and sentences. Establishing links between corresponding segments is called alignment . In particular, linking corresponding sentences is called *sentence alignment* . Such an alignment essentially creates segmentally searchable parallel corpora of collections of documents and their translations.

Sentence alignment is a well established task which does not exclusively refer to 1-to-1 alignments. Sentence boundaries may vary in different translations. However, it usually assumes that information at the sentence level is expressed in the same order in the original document as in its translations. With this assumption, sentence alignment can be modeled as a monotonic mapping process, i.e. an alignment without crossing links. A sample of a sentence aligned bitext is given in figure 2.2.

1:1	I didn't know what to say.	Jag visste inte vad jag skulle säga.
2:3	Her brother said to her, "Why does Ras always say 'longwedge' for -'language', he talks about African 'longwedges'? Sounds so funny."	Brodern inföll: "Hur kommer det sig att Ras alltid säger 'sprak' i stället för 'språk'? Han talar om afrikanska 'sprak', det låter så roligt."
2:1	"Go to hell." Emmanuelle sat up straight.	"Dra åt skogen!" sade Emmanuelle och satte sig kapprak.

Figure 2.2: Sentence alignment from *Nadine Gordimer: "A Guest of Honour"* (aligned at the Department of Computer and Information Science, Linköping University [Mer99b])

Several approaches to automatic sentence alignment have been proposed. The main approaches apply either length based models using correlations between the lengths of corresponding sentences, dictionary based models using correspondences between words and other lexical units, or combinations

[2] http://nl.ijs.si/ME/CD/docs/1984.html
[3] http://www.isi.edu/ koehn/publications/europarl/
[4] http://logos.uio.no/opus/

of both. Additionally, information about the document structure can be used [TRA03] to identify corresponding segments.

Length-based sentence alignment was introduced in [GC91b]. The authors found a significant correlation between character lengths of corresponding sentences for several language pairs. They proposed an alignment model based on a statistical distance measure and a dynamic programming approach. The length-based approach has been applied to a large variety of language pairs (see e.g., [Wu94, TKS96, TN03]) and has proven to be highly accurate for most of them. Some researchers applied sentence lengths in terms of words instead of characters for sentence alignment [BLM91]. However, in [GC91b], the authors demonstrated experimentally that character based models are superior to word based models.

Sentence alignment using lexical information was introduced in [KR88]. There, selected words with similar distributions serve as anchor words for establishing sentence alignments. A geometric approach to sentence alignment using such points of correspondence was proposed in [Mel96b]. In this study, the author introduces an algorithm for finding geometric patterns in a bitext mapped on a 2-dimensional bitext space . Chains of corresponding points in this bitext space are used to align sentences in the parallel corpus. This method has been successfully ported to other language pairs [Mel97b].

Combining lexical information with length-based sentence alignment has been suggested by several researchers (see e.g., [SFI92, JH94]). Various techniques have been proposed for finding corresponding words that may serve as anchor points in a parallel corpus. String similarity measures can be used to find possible cognates [SFI92]. Machine-readable dictionaries can also be utilized for identifying corresponding words [Mel96b]. Distributional models find anchor points by relating word occurrence frequencies. In these approaches, corresponding words are found using measures such as point-wise mutual information[5] in combination with t-score filters [FC94] or using similarity measures between so-called recency vectors of word occurrences

[5]Point-wise mutual information differs from the standard measure of mutual information in information theory. Mutual information $I(X;Y)$ measures how well one random variable predicts another one; i.e. how much information about a random variable Y is included in another random variable X and vice versa. It is defined as the weighted sum of possible event-combinations $I(X;Y) = \sum_x \sum_y p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)}$. Point-wise mutual information considers only one specific "point" of the probability distribution [MS99]. The random variables involved here are binary, i.e. their distribution includes only two probabilities, one that a certain event occurs (e.g. a word occurs in a corpus) and the other that the event does not occur. In this case, point-wise mutual information considers only the point where the event (or the joint event) actually happens and discards the other combinations $I(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)}$. Point-wise mutual information is sometimes referred to as *specific mutual information* whereas the mutual information from information theory is called *average mutual information* [SMH96]. In computational linguistics, the term mutual information has often been used to denote point-wise mutual information. The reader should be aware of this fact when referring to the literature.

[FM94]. In the first approach, *K-vec*, each half of the bitext (source and target language) is split into a number of equally long segments. Frequencies are counted for word pairs that co-occur in corresponding segments, which is the basis for calculating an association measure such as point-wise mutual information. The second approach (*DK-vec*) applies a pattern matching technique called *dynamic time warping* for comparing so-called *recency vectors* from both halves of the bitext. These vectors contain word position distances and are used to describe the "distributional signals" of words occurring in the text. Both techniques are used to identify corresponding words that can be utilized as anchor words in sentence alignment.

Further enhancements and combinations of sentence alignment techniques can be found in the literature, see e.g., [SFI92]. The use of more than two languages is explored in [Sim99]. Automatic sentence alignment is known as a task that can be accomplished with high accuracy, above 90%. However, improvements are still possible in the most difficult cases, especially in connection with "noisy corpora" including divergent and incomplete translations.

2.3 Word alignment

An important implicit resource in parallel corpora is the huge number of translational relations between words and phrases included in the documents. However, such relations cannot easily be defined as monotonic mappings in a bitext space as in the case of sentence alignment. Word order is in general not identical for most language pairs, and boundaries of lexical units are not as easy to detect as sentence boundaries. There is no consistent correlation between the character lengths of corresponding words, at least not for most language pairs. Furthermore, the notion of lexical equivalence presumes that there are identical lexical concepts in both languages and that they behave identically in the given context. This is clearly not always the case. However, natural languages must be compositional in some sense for translation to be possible at all [Isa92]. Thus, many translation relations between these compositional components, i.e. words and phrases, can be found in documents and their translations. Linking corresponding words and phrases in parallel corpora is usually called *word alignment*, a process which can be used as the basis for the extraction of bilingual lexicons.

The type of relation between words varies in parallel texts. Furthermore, the strategy of aligning words and phrases in parallel corpora depends on the task to be accomplished. Usually, word alignment aims at a complete alignment of all lexical items in the corpus, i.e. the goal is to break each bitext segment into sets of corresponding lexical items. This often leads to "fuzzy" translation relations between certain words [MAA02, Vér98, ON00a]

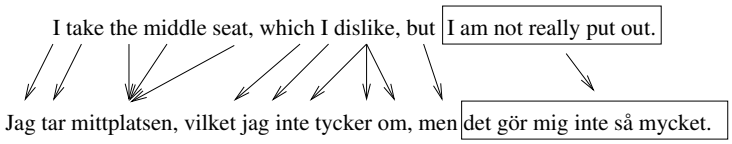


Figure 2.3: An illustrative word alignment example
(from *Saul Bellow: "To Jerusalem and back: a personal account"*)

due to lexical differences, structural and grammatical differences, paraphrased translations, spelling errors, and other divergent translations. The degree of correspondence can be expressed in terms of alignment probabilities, which is useful for many tasks, e.g. statistical machine translation. Bilingual lexicon extraction aims at the identification of lexical word type links in parallel corpora. These links can be inferred from word alignments.

There are generally two approaches to word alignment, the *association approach* using measures of correspondence of some kind, and the *estimation approach* using probabilistic translation models. Association approaches are also referred to as *heuristic approaches* [ON03] or *hypothesis testing approaches* [Hie98]. Estimation approaches are often called *statistical alignment*, e.g. in [ON00a]. Both approaches use some kind of statistics. Hence, I will use the terms estimation and association in order to avoid any confusion between them. In the next two sections, both approaches are briefly introduced and discussed. Some alternative alignment models and a description of evaluation techniques are given thereafter.

2.3.1 Association approaches

Association approaches to word alignment originate mainly from early studies on lexical analysis of parallel data. Lexicographers investigated the use of parallel corpora for the creation of bilingual lexicons and for the support of cross-lingual lexical disambiguation. The relation between word alignment and lexicon extraction is discussed in the following section. Following this, common association measures and resources for such an extraction are briefly introduced, and finally, approaches for handling multi-word units are described.

Alignment and the extraction of bilingual lexicons

The task of bilingual lexicon extraction differs from full-text word alignment in so far as identified translation relations will be used outside their context. Grammatical functions, uncertain relations, and translation irregularities are most likely to be excluded from an extracted lexicon. Furthermore, in a lexicon

extraction task it is not necessary to align all occurrences of a lexical item to corresponding items in the translation. The important part is to find at least one instance of each translation relation to be included in the resulting lexicon; inflectional variants can usually be inferred from other forms. Sometimes, only a sub-set of lexical items has to be considered, e.g. domain-specific terminology.

In general, the following steps have to be taken for aligning bitexts using an association approach:

lexical segmentation: Boundaries of lexical items have to be identified for both languages[6].

correspondence: Possible translation relations between lexical items have to be identified according to some correspondence criteria. This usually results in a collection of weighted word type links, i.e. a translation dictionary with association scores attached to its entries. Contextual features may be attached to this *association dictionary*.

alignment and extraction: The most reliable translations according to the association dictionary are marked in the bitext (alignment). This is commonly done in a "greedy" way using simple search strategies such as a "best first" search in combination with some linguistic/heuristic constraints. A bilingual translation dictionary can be compiled from the aligned items (extraction), which is usually "cleaner" than the previously produced association dictionary.

In spite of the fact that sentence and word alignment are very different from each other, many ideas from research on sentence alignment can be applied to word alignment as well. Previously, several techniques of anchor based sentence alignment have been discussed. Finding anchor points, which

[6]Note that lexical items may refer to single words as well as to phrases or even whole sentence fragments. Note also that it might be necessary to change lexical boundaries for different language pairs. This is often the case when the segmentation into lexical concepts differs between languages. For example, a large concept may be bound to one particular word in one language but in a second language it may be required to use a whole phrase in order to explain the same meaning. However, a third language may use a set of sub-concepts similar to the ones in language two. In this case, lexical boundaries should probably differ when aligning words of language two and three compared with an alignment of words of language one with words of one of the other two languages. Similar problems appear with morphological and derivational differences between languages. For example, in one of our parallel corpora the Swedish compound "regeringsförklaring" is translated into the English noun phrase "statement of government policy" and into the French "declaration de politique générale du gouvernement". An English-French word alignment with links between (statement - declaration), (of - de), (government policy - politique générale du gouvernement) is totally acceptable whereas a Swedish-English alignment requires a link between the Swedish compound and the complete noun phrase in English (similarly for Swedish-French).

is usually defined as the identification of corresponding words, is in fact nothing other than aligning words. Thus, early work on extracting bilingual lexicons was often based on lexical approaches to sentence and paragraph alignment in combination with empirical analysis of lexical co-occurrence, as, for example, reported in [CGHH91]. The *K-vec* algorithm is one of the early techniques for extracting lists of corresponding words using point-wise mutual information scores and t-score filters [FC94]. Similarly, the *DK-vec* algorithm based on similarity measures between word recency vectors can be applied for extracting corresponding words from parallel corpora [FM94]. The *word_align* program [DCG93], which is based on the *char_align* aligner [Chu93], was developed for identifying technical terms and their translations in parallel texts using similarities between character N-grams. It was used in the semi-automatic extraction tool *Termight* [DC94]. Melamed [Mel95] used a cascade of heuristic filters for inducing translation lexicons. He applied string similarity measures and word order heuristics among other things. The same author introduced the notion of *competitive linking* for one-to-one word alignments [Mel96a] which brings up the idea of iterative size reduction mentioned in [Tie97].

Co-occurrence measures

Many word alignment approaches presume sentence aligned corpora. Sentence alignment, as discussed in the previous section, is a reasonably reliable task providing aligned regions that can be used to count frequencies of word pairs co-occurring in these regions. Such co-occurrence frequencies can be used in association measures for the identification of word correspondences.

A common idea behind statistical association measures is to test if two words co-occur significantly more often than it would be expected if they would co-occur purely by chance. The *t-score* is an example of such a test metric. It is derived from the *t-test*, a common hypothesis test in statistics. The general form of this test is the following: $t = (\overline{X_o} - X_h)/SE(\overline{X_o})$ where $\overline{X_o}$ is the mean of the observed values, X_h the expected value according to the hypothesis, and $SE(\overline{X_o})$ the standard error of the observations. Using the central limit theorem, the standard error is defined as the sample deviation (SD) divided by the square-root of the number of experiments K: $SE(\overline{X_o}) = SD(\overline{X_o})/\sqrt{K}$. The t-value gives the distance from the mean of *Student's t-distribution*[7]. The value of t is associated with a *p-value*, the probability mass of the distribution outside the interval from the mean to t[8]. The p-

[7]The t-distribution is used instead of the *normal distribution* for hypothesis tests on random variables with unknown standard deviations. Student's t-distributions depend on the number of observations which determine the *degree of freedom*. The distribution approaches the standard normal distribution for high degrees of freedom.

[8]One distinguishes between one-tail and two-tail tests depending on whether the hypothesis is directional or not.

value corresponds to the statistical significance of the difference between observation and hypothesis. A low p-value indicates statistical evidence for rejecting the hypothesis. Now, the t-test can be applied as an association measure between translated word pairs as follows: The joint probability of co-occurring words can be observed in a parallel corpus ($\overline{X_o} = p(w_s, w_t)$). Using the hypothesis that both words co-occur purely by chance, i.e. the distributions of both words are independent of each other ($X_h = p(w_s)p(w_t)$), we can apply the t-test to find out if there is any statistical evidence for rejecting this hypothesis, i.e. evidence for a dependence between w_s and w_t. The t-test is often used as an association measure because its value becomes larger when there is stronger evidence for rejecting the independence hypothesis. This measure is usually called the *t-score*. In the case of bilingual co-occurrence, the standard error is estimated as defined above, where K is the number of aligned bitext segments and the sample deviation is approximated using the square root of the observation mean[9] ($SD(\overline{X_o}) = \sqrt{p(w_s, w_t)}$) [FC94]. Probabilities and standard deviations are estimated from the corpus using relative frequencies.

$$t \approx \frac{p(w_s, w_t) - p(w_s)p(w_t)}{\sqrt{\frac{1}{K}p(w_s, w_t)}} \tag{2.1}$$

Another association measure based on co-occurrence is the *Dice coefficient*. This coefficient can be used to measure the correlation between two events (the occurrences of w_s and w_t) as follows:

$$Dice(w_s, w_t) = \frac{2 * p(w_s, w_t)}{p(w_s) + p(w_t)} = 2 * \left(\frac{1}{p(w_s|w_t)} + \frac{1}{p(w_t|w_s)} \right)^{-1} \tag{2.2}$$

The formula above shows that the Dice coefficient is in fact the *harmonic mean* of the two conditional probabilities $p(w_s|w_t)$ and $p(w_t|w_s)$. It therefore produces values between 0 and 1, where 1 refers to the strongest correspondence.

A third statistical association measure is point-wise mutual information, derived from information theory. Mutual information $I(X;Y)$ measures the amount of information common to two random variables X and Y. It is defined as the difference between the *entropy* $H(X)$ of one variable and the entropy $H(Y|X)$ of another variable given the first one. Entropy is a measure of the information content of a random variable: $H(X) = -\sum_x p(x) log_2 p(x)$. Using

[9] The random process of generating bigrams is modeled as a Bernoulli trial with $p = p(w_s, w_t)$ for the probability of the bigram $w_s w_t$ to be produced and $(1-p)$ for the probability of any other outcome. Variances of such distributions can be approximated as $\sigma^2 = p(1-p) \approx p$ if p is small, which is the case for most bigrams in a corpus [MS99].

this definition, mutual information is calculated as follows:
$I(X;Y) = H(X) - H(Y|X) = \sum_{x,y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)}$. Now, we can assume that words that have a lot of information in common are likely to be mutual translations. Applying mutual information, we can set the following parameters: X is a random variable that produces the events w_s (the word w_s occurs) or $\neg w_s$ (the word w_s does not occur). Y is a random variable that produces w_t and $\neg w_t$. The joint probability $p(w_s, w_t)$ describes the probability of w_s and w_t to co-occur in the bitext. $p(w_s)$ and $p(w_t)$ are the probabilities of w_s and w_t to occur in the corpus, respectively. All probabilities can be estimated from the corpus. *Point-wise* mutual information considers only "one point" in the distributions of X and Y, namely $p(w_s) = p(X = w_s)$ and $p(w_t) = p(X = w_t)$. Hence, the definition of point-wise mutual information for co-occurrence is as follows:

$$I(w_s, w_t) = log_2 \frac{p(w_s, w_t)}{p(w_s)p(w_t)} \tag{2.3}$$

In this way, point-wise mutual information "measures the reduction of uncertainty about the occurrence of one word when we are told about the occurrence of the other" [MS99, p. 183].

Many other measures of correspondence have been applied to parallel corpora with slightly different results. Two examples are the Φ coefficient [GC91a], and the log-likelihood measure [TB02]. Advantages and weaknesses of specific measures have been discussed much in the literature (see e.g., [CG91, Dun93, SMH96, MS99]). A comparison of association measures can be found in [RLM00]. We restrict this description to the three measures introduced above, i.e. the t-score, the Dice coefficient and point-wise mutual information, as they are used in our word alignment approaches.

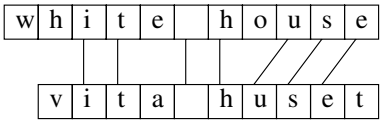
String similarity measures

Other empirical alignment techniques than the ones mentioned above are based on measures of string similarity. Many *cognates* can be found especially in bitexts of closely related languages. The term cognate denotes here etymologically related words across languages. Many cognates can be identified by way of their spelling. Simple string matching algorithms can be used to exploit this property. Initial character sequences are one simple form of cognate identification [SFI92]. Variants of the Dice coefficient can also be used to compare common character n-grams in order to find a level of similarity between strings [BM96]. For example, using character bigrams, the Dice coefficient for string similarity can be formulated as follows:

$$Dice = \frac{2 * |bigrams(x) \cap bigrams(y)|}{|bigrams(x)| + |bigrams(y)|} \tag{2.4}$$

Another approximate string matching algorithm is the *longest common subsequence ratio* (LCSR) as described, for instance, in [Mel95]. LCSR is defined as the ratio of the longest common sub-sequence (LCS) of characters of two strings and the length of the longest of both strings. LCSs can be found using dynamic programming approaches for arbitrary pairs of strings [Ste92]. LCSR is a measure with values between 0 (completely distinct, i.e. LCS=0) and 1 (identical). An example is given in figure 2.4.

$Dice(\text{'white house'}, \text{'vita huset'}) = \frac{2*|\{\text{'it'}, \text{' h'}, \text{'us'}, \text{'se'}\}|}{10+9} = \frac{8}{19} \approx 0.42$



$LCSR(\text{'white house'}, \text{'vita huset'}) = \frac{|\text{it huse}|}{max(|\text{white house}|, |\text{vita huset}|)} = \frac{7}{11} \approx 0.64$

Figure 2.4: String similarity measures.

Note that string matching algorithms are just a tool for finding possible cognates using the assumption that cognates are similar in spelling. This presumes a similar alphabet and similar spellings of etymologically related words. Algorithms for the construction of weighted string similarity measures that include mappings between non-identical characters are described in [Tie99a] and in section 5.1.1 on page 50. Furthermore, the problem of false friends is common and is usually faced with string length thresholds (e.g. ≥ 4 characters). A comparison of string matching techniques and linguistically motivated methods for cognate extraction has been presented in [Bor98].

External alignment resources

There are many external resources that can be employed in word alignment, for instance, collections of machine-readable bilingual dictionaries (MRBD). External resources define common relations between words and phrases that can be used in word alignment. The use of bilingual dictionaries is, for instance, discussed in [ON03] and [Mel95]. The impact of such resources on the performance of automatic word alignment depends very much on their size and appropriateness with respect to the corpus and its domain.

Another type of external alignment resource is language-specific expert knowledge, which can be put into alignment systems by means of heuristics.

Word order constraints and position relations are commonly used for boosting word alignment. Such constraints can be expressed in terms of monotonicity assumptions [OTN99] or in terms of position weights [AMA98]. Relations between words of similar word classes (using, e.g., part-of-speech labels) can also be used as pre-defined heuristics. This is, for example, applied in the alignment experiments presented in [TB02] and in the lexicon extraction approach in [Mel95]. Similarly, syntactic relations can be integrated in word alignment systems. Morphosyntactic information and syntactic function are used, for instance, in the interactive word aligner I*Link [AMA02]. The combination of such resources with statistical alignment techniques is usually not straightforward. The adjustment of parameters is a common problem in word alignment. Section 5.1.2 describes a sequential combination of alignment techniques and resources, and section 5.1.3 presents a probabilistic approach.

Multi-word units

The one-to-one alignment assumption for word alignment is insufficient for most language pairs, as already mentioned in the introductory part of section 2.3. Several studies have been made on the integration of so-called *multi-word units* (MWUs) in word alignment. By MWUs we refer to word sequences and word groups, which express structural and conceptual units such as complex noun phrases, phrasal verbs, idiomatic expressions, and other phrasal constructions that should not be split up in an alignment process.
Two general techniques are used for dealing with MWUs: prior identification of collocations and dynamic construction of MWUs during the alignment process. Exhaustive research on the identification of collocations has been carried out in studies of monolingual terminology extraction. Statistical association measures have been proposed for finding MWUs. The use of such measures for monolingual lexical analysis has been presented in, e.g., [SM90, CGHH91, Dun93, MNA94]. Another way of identifying complex terms is to use linguistic knowledge such as part-of-speech information or phrase-structure analyzes. Typically, noun phrases are emphasized in work on automatic multi-word term extraction [JK95, Arp95]. A common way of noun phrase identification is to match language-specific part-of-speech tag patterns. Hybrid systems using statistical as well as linguistic information have been proposed, e.g., in [Dai95, MA00]. Parsing techniques can also be applied for finding relevant MWUs. In particular, shallow parsing techniques have recently been developed for shallow analyzes of unrestricted texts [Abn91]. Shallow parsing can be modeled as a classification task [RM95] and has become very popular in the machine-learning community (see e.g., [Rat98, FHN00, KM01, Meg02, TKS02]).

The use of structural information in bilingual term extraction has been investigated in a number of studies. In [vdE93], the author argued that noun

phrases represent a better level to compare than words for the alignment for Dutch and English. He used a simple pattern-based extraction of noun phrases and association measures for selecting phrase translations. Other studies use statistical or hybrid approaches for the identification of phrasal units in bitext segments before proceeding with word alignment. Ahrenberg et al [AMA98] apply an n-gram based phrase retrieval tool in their word alignment system in order to identify recurrent MWUs in both languages of each bitext segment. They improved the system by adding language filters using classified function word lists and entropy thresholds to their frequency-based phrase extractor [MA00]. A similar approach has been applied in [Tie99c] using point-wise mutual information for the identification of MWUs prior to alignment. Alternatively, such units may be searched for "on-the-fly" during the alignment process. Such an approach has been introduced in [SMH96]. The authors produce MWU alignments using a statistical collocation extractor [SM90] for the source language and an iterative alignment procedure for the identification of possible translations in the target language. The system starts the alignment of collocations with a link to single words and compares the association score iteratively with scores for links to larger units. Another "dynamic" segmentation approach has been used in the experiments presented in [Tie99c]. In this study, experiments using an iterative segment expansion procedure for both the source language and the target language have been carried out and compared to alignment results with prior MWU detection. This dynamic segmentation approach favors MWU alignments in cases where their association is stronger than associations between parts of the MWUs. Melamed [Mel97a] investigated another iterative approach for the automatic discovery of non-compositional compounds (as one type of MWUs) in parallel corpora. In his algorithm, alignments including MWU candidates are compared to alignments without them. He uses mutual information scores as the "objective function" for the "net gain" when including the candidate MWUs. Candidate MWUs are adjacent words excluding previously defined stop words. This technique can be used for both languages by swapping the direction of the alignment. This, finally, results in a translation model for *bag-of-word alignments* [Mel00].

2.3.2 Estimation approaches

By estimation approaches we refer to word alignment using probabilistic alignment models that are estimated from parallel corpora. Most work in this field has been inspired by the work on statistical machine translation introduced in [BCD$^+$90]. In the next sections, I will briefly review the principles of statistical machine translation and describe the application of such models to word alignment and lexicon extraction tasks.

Statistical machine translation (SMT)

SMT is an application of the noisy channel model from information theory [Sha48] to the task of machine translation. The source language S and the target language T are considered to be random variables that produce strings such as sentences. Translation is modeled as a transmission of a source language string[10] $\mathbf{s}$ through a noisy channel that transforms it into a string $\mathbf{t}$ in the target language. The probability $P(\mathbf{t}|\mathbf{s})$ is then interpreted as the probability of $\mathbf{t}$ being a proper translation of $\mathbf{s}$. In a noisy channel model the target string ($\mathbf{t}$) is considered to be the observable part of the system and the task of the model is to find the original input string ($\mathbf{s}$) that has been transmitted through the channel in order to produce $\mathbf{t}$. Using Bayes' rule the probability of the input string $\mathbf{s}$ given the observation $\mathbf{t}$ is defined as follows:

$$P(\mathbf{s}|\mathbf{t}) = \frac{P(\mathbf{t}|\mathbf{s})P(\mathbf{s})}{P(\mathbf{t})} \tag{2.5}$$

According to the model, the most likely solution can be found by using the $argmax_{\mathbf{s}}$ function, which returns the argument $\widehat{\mathbf{s}}$ out of all possible values for $\mathbf{s}$ that maximizes the given function:

$$\widehat{\mathbf{s}} = argmax_{\mathbf{s}} P(\mathbf{s}|\mathbf{t}) = argmax_{\mathbf{s}} \frac{P(\mathbf{t}|\mathbf{s})P(\mathbf{s})}{P(\mathbf{t})} \tag{2.6}$$

Due to the fact that $P(\mathbf{t})$ is independent of $\mathbf{s}$ and, consequently, is constant for all possible strings $\mathbf{s}$; it can be ignored in the maximization procedure. The fundamental equation of the statistical machine translation model is therefore expressed as the following search problem:

$$\widehat{\mathbf{s}} = argmax_{\mathbf{s}} P(\mathbf{t}|\mathbf{s})P(\mathbf{s}) \tag{2.7}$$

$P(\mathbf{s})$ is called the language model and $P(\mathbf{t}|\mathbf{s})$ the translation model, which is to be estimated from sentence-aligned parallel corpora. However, estimating $P(\mathbf{t}|\mathbf{s})$ directly from a corpus is impossible because of the sparse data problem. The majority of segments in any parallel corpus, let it be as big as possible, will be unique. Even worse, most of the possible sentences $\mathbf{s}$ and $\mathbf{t}$ of two languages will not occur in any training corpus and therefore according parameters for a translation model are impossible to estimate. Consequently, the translation

[10] Boldface variables such as $\mathbf{s}$ denote strings of outcomes of a random variable such as the source language S. Probabilities such as $P(\mathbf{s})$ denote probabilities of events $\{S = \mathbf{s}\}$, i.e. $P(\mathbf{s})$ is a short form for $P(S = \mathbf{s})$.

model has to be decomposed into distributions of smaller units, which recur more frequently in the training data and are more likely to appear again in unseen data. Most decompositions are based on the translation models proposed in [BDDM93]. The first step in decomposing the general translation model is to introduce another random variable A denoting the alignment between sub-strings (i.e. words) of the source and the target language strings. Using all possible alignments $\mathbf{a}$ between $\mathbf{s}$ and $\mathbf{t}$, the translation model can be re-written as follows:

$$P(\mathbf{t}|\mathbf{s}) = \sum_{\mathbf{a}} P(\mathbf{t}, \mathbf{a}|\mathbf{s}) \tag{2.8}$$

The alignment in SMT is usually modeled as a sequence of *hidden* connections between words in the target language string and words in the source language string. More specifically, each word in the target language string $\mathbf{t}$ is connected to exactly one word in the source language string $\mathbf{s}$, which can be expressed as a natural number representing the position of the connected source language word in the sentence. In order to handle words that do not have a possible equivalent in the other language, a special *empty word* is introduced at position 0 in the source language string. Considering the string $\mathbf{s} = s^L = s_0 s_1 s_2 .. s_L$ of L source language words (plus the empty word s_0) and the translation $\mathbf{t} = t^M = t_1 t_2 .. t_M$ of M target language words, an alignment is a sequence of M connections in the form $\mathbf{a} = a^M = a_1 a_2 .. a_M$ with $a_m \in \{0, .., L\}$. This alignment model is also called a *directional alignment model* because it is not symmetric. It disallows multiple connections from one target language token to several source language tokens. However, the authors of [BDDM93] argue that this problem can be overcome by using (possibly overlapping) multi-word units, which they call *cepts*, in order to find appropriate alignments. This idea is similar to the principle of prior identification of collocations as discussed in section 2.3.1.

Using the directional definition of word alignment, the translation model in equation 2.8 can be decomposed into the following form:

$$P(t^M|s^L) = \sum_{a^M} P(t^M, a^M|s^L) = P(M|s^L) \sum_{a^M} \prod_{m=1}^{M} P(t_m, a_m|t^{m-1}, a^{m-1}, s^L) \tag{2.9}$$

In other words, the joint probability of the target language string and its alignment sequence, given the source language string, can be expressed as the product of the probabilities of all target language words t_m and their alignments a_m, given the previous words t^{m-1} and their alignments a^{m-1} and given the

source language string s^L. The sum of all joint probabilities is multiplied with the probability of the length of the target language sentence, given the source language string $P(M|s^L)$. In this way, the translation probability has been decomposed into one parameter per target language word for each possible alignment position and alignment context. It is still not feasible to estimates these parameters directly from corpora due to the large number of dependencies in the parameters, which still cause a large data sparseness problem. In SMT research, approximations of the translation model above are applied using different independence assumptions.

Most translation models that have been used in the SMT community are based on the five models introduced in [BDDM93] by researchers at IBM[11]. All their models build on the directional word-to-word alignment model discussed previously. The idea behind their five-model-scheme is to start with a very simple model before progressing to more complex ones. The output of simpler models can be used in this way to initialize the following models.

Model 1 is a simple word translation model, which basically makes use of co-occurrence of corresponding words in sentence aligned bitext segments. It is initialized with a uniform distribution of word translation probabilities. Model 2 adds local dependencies by introducing position parameters (*distortion*) to the translation model. In model 3, so-called fertility parameters are introduced. Fertility parameters represent the probability of words to be aligned to a certain number of corresponding words. Modeling fertility copes with the fact that certain words tend to be aligned to multiple words and other words do not. Using fertility probabilities, multiple connections to one word can be penalized or supported. Model 4 includes additional dependencies on the previous alignment and on the word classes of surrounding words in order to handle MWUs, which tend to stick together. Word classes can be learned automatically from bilingual corpora using clustering techniques [Och99]. Model 5, finally, gets rid of the deficiency problem of models 3 and 4. Deficiency of these models means that parts of the probability distributions are reserved for impossible events such as alignments to word positions outside of the sentence boundaries. However, removing deficiency is a rather expensive task and complicates the model. Och and Ney [ON00b] present adjustments for handling the deficient models, which makes it possible to skip the computation of model 5 as it does not provide any significant improvement.

Several variants of the IBM models have been proposed in the literature. One way to model bilingual alignment is to use hidden Markov models (HMMs) as described in [VNT96]. IBM's translation models 1, 2 and 3 can be formalized in a zero-order HMM (where model 3 has additional

[11] Translation models from this study are often referred to as the IBM models 1 to 5.

fertility parameters) and model 4 can be expressed as a first-order HMM with additional fertility parameters [ON00a]. The HMM based translation model is decomposed into an *alignment model* using a chain of hidden alignment parameters $P(a_m|a_{m-1})$ and a *lexical model* with lexical translation probabilities $P(t_m|s_{a_m})$. Statistical translation models can be improved using dependencies on word classes [TIM02], smoothing techniques for the estimation of probabilities [ON03], and external dictionaries [ON00b, ON00a]. Additional contextual dependencies can be integrated into the lexical probabilities using log-linear models and a maximum entropy approach for training [BDD96, GVONC01]. Investigations on adding string similarity measures for cognate identification [GKK03] and on integrating syntactic information [CKY03] have recently been carried out.

Another problem that has to be addressed in statistical alignment models is the handling of multi-word units (MWUs) in both languages. Directional alignment models allow the connection of source language words with multiple target language words but not vice versa. Alignment templates have been introduced in [OTN99] in order to handle MWUs in a more symmetric way. Alignment templates are bilingual sets of word classes that are internally linked using a two-dimensional alignment matrix. These templates are used to link sequences of source language words with sequences of target language words, i.e. to perform a phrase level alignment [VOT$^+$00].

Translation models are trained using the expectation-maximization algorithm (EM), which is an iterative optimization strategy for approaching a local maximum of the likelihood of given training data according to the parameters of a probabilistic model. EM is useful if there are "hidden" parameters which cannot be estimated directly from data. Alignment probabilities are typical examples of such parameters because links between words are not present in the training data. EM starts with an initial guess for all free parameters in the model and updates them iteratively by maximizing the likelihood function until the process converges at a local maximum[12].

A translation model can be used together with a monolingual language model for S (for instance based on n-grams) to translate unseen sentences from language T to language S using the estimated parameters and equation 2.7. A translation model can also be used to find the most likely alignment between words in the training data according to the model. This alignment is called the *Viterbi alignment* and can be used to extract bilingual lexical data from the bitext.

[12]For efficiency reasons, approximate estimation techniques have to be used when running EM on fertility based models.

2.3.3 Alternative alignment models

There are alternative proposals for modeling translation relations statistically besides the IBM models that were described above.

The author of [Kup93] presents a method for mapping noun phrases using an iterative estimation approach based on the EM algorithm. He employs part-of-speech taggers and noun phrase recognizers for both languages in a bitext, and re-estimates the joint probability of source language noun phrases and target language noun phrases at certain positions in the bitext using EM.

Gaussier [Gau98] describes an approach using alignment graphs, which he calls *alignment flow networks* . An alignment network describes directional word-to-word connections as edges with attached flow costs between nodes in the graph. A network has a source node and a sink node and the best alignment is defined as the path through the network for which the total cost flow is minimal. Costs are defined as inverse association probabilities, i.e. a minimal cost flow is defined as the connection for which the association probability is at its maximum. Probabilities of word connections (association probabilities) are modeled as two independent probabilities, the one of linking two positions and the one of linking two words. Association probabilities are assumed to be independent of each other, which leads to the following model of the joint probability[13]:

$$P(a, s_L, t_M) = \prod_{i=1}^{L+M} P(s_i, t_{a_i} | a^{i-1}) \tag{2.10}$$

The alignment flow network model is trained using an approximate EM algorithm. The system has been applied to the task of bilingual terminology extraction. Fertility graphs can be included to handle multi-word terms. The general procedure is similar to the ideas of Smadja [SMH96]: First, candidate terms are identified for one language. Secondly, possible translations of these terms are identified using the flow network model.

Another estimation approach to word alignment is described in [Hie98]. The author uses a two-dimensional contingency table for representing translation frequencies between word types of each language in a given bitext. The free parameters of the model are the probabilities that are connected with each table cell, i.e. probabilities that the words, which correspond to the table cell, are translations of one another in the corpus. Translation pairs are assumed to be independent of each other, which makes the probability estimations a function of the translation frequencies in the table. Now, the free parameters are estimated using the EM algorithm . First, the cells in the contingency

[13] The notation follows the one which has been used in section 2.3.2.

table are filled with initial estimates of the probability parameters. Secondly, the expected cell frequencies are calculated for each aligned sentence pair in the corpus using the observed words and the current probability parameters (E-step). Then, new probability estimates are calculated using the maximum likelihood estimator (M-step). Finally, the E-step and the M-step are repeated until the parameters converge at a local maximum. Two different variants of the model are introduced in the paper, one with an explicit one-to-one word assumption and a second model with the possibility of many-to-one alignments. The second model is more flexible. However, it runs into maximization problems because of alignments of different lengths. Both models have a large search space which causes efficiency problems in the training process. Therefore, approximate EM techniques are used to approach the local maximum. The author also ran experiments with pre-processing steps such as compound splitting, which lead to a better performance than both models without pre-proccessing.

Another word-to-word alignment model has been introduced by Melamed [Mel97c]. This model is a mixture of a standard association approach and a statistical estimation approach. The author initially applies the *competitive linking algorithm*, as already mentioned in section 2.3.1, using log-likelihood tests as a measure of association. Melamed introduces two hidden parameters, which have to be learned from the data. One parameter λ^+ represents "true positives", i.e. the probability of a link given that co-occurring word pairs are mutual translations of each other. The other parameter λ^- represents "false positives", i.e. the probability of links given that co-occurring word pairs are not mutual translations. Using these parameters, the likelihood of two word types being mutual translations is defined as the ratio of the probabilities of two words being linked, given the co-occurrence frequency and the positive parameter λ^+, and the probability of the same two words being linked, given the co-occurrence frequency and the negative parameter λ^-. In other words, the model for linking word types depends on the co-occurrence frequencies and the two hidden parameters which have to be estimated in a maximization procedure. This can also be called a *re-estimation of association measures* using false and true positives as hidden parameters.

2.3.4 Evaluation

The reader might have observed that evaluation has not been mentioned in the previous sections, and no comparison in terms of performance has been made. Evaluation of alignment is a tricky part. Most studies refer to recall and precision measures, which have been derived from information retrieval. Precision (P), giving the number of correctly aligned items ($|correct \cap aligned|$) in proportion to the number of obtained items ($|aligned|$),

and recall (R), giving the number of correct results in proportion to the number of correct items in total ($|correct|$), seem to be reasonable measures for comparing system performances. A balanced F-value[14] is often used to combine both measures for a comparison of the overall performance:

$$P = \frac{|aligned \cap correct|}{|aligned|}, R = \frac{|aligned \cap correct|}{|correct|}, F = \frac{2 * P * R}{P + R}$$

However, precision and recall values are not as straightforward to estimate in the case of alignment as in information retrieval. Usually, results are not easily judgeable as completely correct or completely wrong. *Partiality* is a common phenomenon in word alignment results. The possibility of MWU links causes the system to return partially correct links in many cases. Link proposals including at least one correct word on both sides of the link (source and target language) are called *partially correct links*. These links are not captured by standard precision and recall. Therefore, the *degree of correctness* has to be integrated in evaluation measures in word alignment. However, this is not as straightforward as one might expect. Several approaches will be discussed later on.

In general, complete correctness of any word alignment cannot be expected. This is due to the nature of translation of natural language. Translations are correspondences in context and word alignment tries to break parallel texts into related units at the word level, which is not always feasible. This has been experienced in several attempts of manually creating word aligned reference material [Mel98, Vér98, Mer99a]. Word alignment also depends on various corpus characteristics. There will be differences depending on genre, language pair and the style of individual translators. Another crucial factor is the purpose of an alignment experiment, either lexicon extraction or full-text word alignment. In the first case, the extracted lexicon is to be evaluated, in the latter case, aligned tokens in the corpus have to be judged. The focus in lexicon extraction is set on content words whereas function words can be neglected. A word alignment system aiming at the creation of aligned bitexts has to evaluate token links within the corpus even for divergent translations[15] and highly ambiguous function words. *Translation spotting* is another type of application that has been studied in an alignment competition [VL00], in which translations of a number of given source language terms are sought in bilingual corpora.

[14] The balanced F-value is derived from the weighted F_β measure, which is defined as the ratio $F_\beta = ((\beta^2 + 1) * P * R)/(\beta^2 * P + R)$. Setting $\beta = 1$ "balances" precision and recall, i.e. both rates are weighted to be equally important.

[15] With divergent translations we refer to insertions, deletions, errors or other unexpected parts in translated text.

Gold standards

Automatic evaluation using a reference alignment (=*gold standard*) is often preferred over manual a posteriori evaluation. The main advantage of reference alignments is their re-usability once they are created. The main difficulty is to produce representative samples of reliable reference alignments.

In some studies sample bitext segments have been completely aligned by hand in order to create gold standards [Mel98, ON00b]. In other studies word samples from the corpus [VL00, AMST00] were used. The word sampling approach has the advantage that the evaluation can be focused on certain word types such as content words or words from certain frequency ranges [Mer99b]. Segment based alignment has the advantage that linking decisions are often easier to make when surrounding context is to be aligned as well. Furthermore, recall and precision measures are more straightforward for completely aligned segments than for a sampled gold standard.

Links between MWUs can be expressed in different ways in gold standards. In [AMST99], MWUs are treated as complex units and links between them are established in the same way as between any other word pair. [Mel98], [ON00b] and [MP03] treat MWUs as sets of words and a link between two MWUs is expressed as the exhaustive set of one-to-one word links between the members of both MWUs. A consequence of splitting links between MWUs into one-to-one word links is that the number of links increases compared to the complex unit approach, which certainly has an impact on the evaluation measures precision and recall. Consider the example in table 2.1.

	Gold standard	
bitext segment	complex alignment	MWU splitting
no one is very patient	no one → ingen	is → visar
ingen visar mycket tålamod	is patient → visar tålamod	is → tålamod
	very → mycket	patient → visar
		patient → tålamod
		no → ingen
		one → ingen
		very → mycket

Table 2.1: Types of gold standards.

Let us assume that an alignment system finds the links "patient → tålamod" and "very → mycket". A restrictive[16] evaluation system would score 1 out of 2 correct links using the complex MWU notation. In the second approach, the same alignment would yield only 2 out of 7 links. Recall differs between 0.5 for the restrictive MWU approach and about 0.29 for the splitting approach.

[16]Restrictive evaluation refers to evaluation disregarding partly correct alignments.

Precision differs between 0.5 and 1. The F-value for the first approach is 0.5 and ca 0.36 for the second approach. The MWU splitting approach has the advantage that it counts the partially correct link "patient → tålamod", which is not captured by the restrictive complex unit approach. However, the scores may be blurred as in the example because of the increasing number of links when splitting MWU links.

Another problem that has to be solved when producing gold standards is the alignment of divergent translations. Previous studies have demonstrated how difficult it is to agree on manual word alignments [Mel98, Vér98, Mer99a]. Approaches to handle uncertain links are quite similar in these studies. [Vér98] uses confidence levels as a degree of certainty on a scale between 0 and 3. [Mer99a] uses "fuzzy" markers for labeling uncertain links. [ON00b] uses the marker 'P' for "probable" alignments.

A last group of links is commonly called *null links* referring to not translated words. Certain words do not have any correspondence in another language such as the auxiliary verb 'do' in English questions or negations. Other words are simply not translated and therefore cannot be aligned.

In some cases it is hard to decide if words should be aligned as fuzzy links, non-aligned null links, or if they should be included in a larger link unit. Therefore, detailed guidelines are necessary for manual annotators when creating gold standards [Mel98, Vér98, Mer99a].

Evaluation metrics

Precision and recall can be defined in several ways according to the gold standard and the representation of MWU links, fuzzy links and null links. The main difference is to be found in the treatment of partially correct links. Partiality is measured in different ways. The evaluation measures of the ARCADE word alignment track were tailored towards the task of translation spotting, i.e. the search for proper translations of given source language terms. Therefore, the measures consider only tokens in the target language. They are defined as follows:

$$R_{arcade} = \frac{1}{X}\sum_{x=1}^{X}\frac{|aligned_{trg}^{x} \cap correct_{trg}^{x}|}{|correct_{trg}^{x}|}, P_{arcade} = \frac{1}{X}\sum_{x=1}^{X}\frac{|aligned_{trg}^{x} \cap correct_{trg}^{x}|}{|aligned_{trg}^{x}|}$$

$aligned_{trg}^{x}$ is the set of target language words found by the alignment system for the source language term in link x of the gold standard. $correct_{trg}^{x}$ is the set of correct target language words of link x in the gold standard. Null responses are counted as links to a special "empty" word, i.e. if the system does not find any link for reference x the set $aligned_{trg}^{x}$ will be set to $\{empty\_word\}$. Consequently, null responses (i.e. missing alignments) have an impact on both, precision and recall.

Normally, word alignment also has to cope with the task of finding the correct source items (words and phrases) to be aligned. Therefore, other measures are needed for general word alignment systems that do not have access to given source language terms. The MWU splitting approach is one way to handle partiality in a symmetric way. In [ON00b], the following metrics have been proposed (henceforth the *SPLIT* measures):

$$R_{split} = \frac{|aligned \cap sure|}{|sure|}, P_{split} = \frac{|aligned \cap probable|}{|aligned|}$$

The set of *aligned* one-to-one word links is compared with the set of *sure* one-to-one word links from the gold standard for measuring recall. Links that have been marked to be *sure* are a sub-set of all one-to-one word links in the gold standard, which are referred to as *probable* links. In contrast to recall, the complete set of (*probable*) links is used for measuring precision as they can be "correctly" proposed by the system. The authors of [ON00b] also suggested a combination of both measures (essentially a complementary F-value) which they call the *alignment error rate*.

$$AER = 1 - \frac{|aligned \cap sure| + |aligned \cap probable|}{|aligned| + |sure|}$$

The *SPLIT*-measures above are designed for word-to-word alignment approaches using the split type of MWU links. Measures for word alignment evaluation, using complex MWU link references, are presented in section 5.2.

2.4 Applications

In the previous sections common techniques for processing parallel data were presented. Several applications of aligned data have been mentioned already. In this section, I will list some additional applications, tools and projects that have been based on parallel corpora and techniques from above. See also the illustration in figure 2.1 on page 7 for the relation between parallel corpora, alignment approaches, and the applications which are mentioned here. Note that this description is not intended as a comprehensive list of tools and projects on this subject.

Sentence aligned parallel corpora are directly applicable for supporting translators in their daily work. Translation memories have been used for a long time by human translators and sentence aligned bitexts can be used as such without any further processing. Extending the functionality of translation memories by aligning even sub-sentential parts leads to the idea of example-

based machine translation (EBMT) [Bro96]. Several techniques have been proposed for generalizing aligned segments [Bro00] and putting "bits and pieces" together that have been derived from old translation examples.

Statistical machine translation was introduced above. SMT systems become ever more popular due to recent improvements of translation models and increased power of today's computer technology. SMT systems have the advantage that they can be developed very fast once there are tools and sufficient training data available for the particular language pair. SMT systems have the disadvantage that they rely on training and the statistical model. Corrections and improvements are hard to integrate in the set of estimated parameters which are usually not human readable. However, SMT has been used in several applications starting with the *Candide* system at IBM [BBP+94] and moving to the VERBMOBILE project on speech translation [VOT+00]. The flexibility of SMT systems has been proven by the "MT in a Day" experiment which was carried out at the NSF Workshop on statistical machine translation at Johns Hopkins University [AOCJ+99]. Many teams work on the improvement of SMT systems. *Co-training* of SMT models using more than two languages is one way to boost the performance of a translation system [CBO03].

Recently, interactive machine translation (IMT) has been studied in connection with statistical translation approaches. The idea of translation predictions for IMT has been suggested in [FIP96] and implemented in the *TransType* system [LFL00]. The SMT framework has been integrated in the system using automatically constructed word hypothesis graphs for the efficient search of possible translation completions [OZN03].

Another obvious application of parallel corpora is the extraction of bilingual terminology. Several systems have been developed using word alignment techniques as described above. *Termight* uses Church's character-based alignment approach *char_align* [DC94], *TransSearch* uses IBM's model 2 [MH96], and *Champollion* uses Smadja's collocation aligner [SMH96]. Terminology extraction techniques have successfully been ported to a variety of language pairs among them less related languages such as English and Japanese [FM97] or English and Chinese [WX94].

Related to terminology extraction is the field of lexicography. The use of bilingual data for building translation dictionaries has been investigated in several projects. *BICORD* is one example of an attempt to combine bitexts and machine-readable dictionaries for building and extending bilingual dictionaries [KT90]. *Dilemma* is another lexicographic tool that re-uses existing translations [KKN+94]. Many more projects aim at the automatic or semi-automatic extraction of bilingual lexicons for different language pairs (see e.g., [RM97, ARM01, AMA02]).

Furthermore, extracted translation dictionaries can be applied to machine

translation as lexical resource in, for instance, direct and transfer based machine translation systems (see e.g. [Ahr99, MR01, SFT+02, ISM03]), or example-based machine translation (see e.g.,[Bro97].

Another field of research where parallel data can help is the field of word sense disambiguation. Ambiguities are distributed differently in natural languages. This fact can be used for cross-lingual comparisons, which may help to disambiguate words and to identify concepts in context [GCY92, DR02]. Dyvik explores translations as semantic mirrors of ambiguous words [Dyv98]. Translation alternatives of a language sign (i.e. word) describe a so-called *t-image* of the sign in this language. t-images can be reversed. i.e. sets of translational correspondences in the original language can be found for all words in the *t-image*. Intersection between these sets in inverse t-images represent conceptual distinctions of the original sign. Dyvik uses multiple inversions and several heuristics for producing semantic networks for ambiguous words in a parallel corpus [Dyv02]. Furthermore, these networks can be linked between the two languages.

A last application of parallel corpora to be mentioned here is the adaptation of language tools to new languages with the help of parallel data. Robust text analysis tools, which exist for one language, can be ported to other languages by projecting analyzes (such as part-of-speech and chunks) from one language to another in a parallel corpus [Bor99, YNW01, Bor02]. Similarly, a third language may be used to induce word alignments between two other languages [Bor00].

2.5 Summary

In this chapter, basic concepts and techniques of the work with parallel corpora have been presented. The compilation of parallel corpora using alignment techniques such as automatic sentence alignment has been introduced. Common methods for the alignment of sentences have been discussed in section 2.2. Word alignment techniques have been described in detail as they are essential for the extraction of lexical correspondences from bilingual parallel corpora. Two main approaches to word alignment have been discussed, which we refer to as *association approaches* and *estimation approaches*. Association measures are widely used for the identification of translational correspondences. Section 2.3.1 describes common measures used in association approaches to word alignment. Their application in our alignment systems is presented in chapter 5. In statistical machine translation, probabilistic alignment models are used to estimate alignment parameters between words in parallel corpora. This approach is referred to as the estimation approach to word alignment in the present thesis. Statistical translation models, as the ones described in section 2.3.2, are mainly studied

for the purpose of machine translation. However, research on word alignment and bilingual lexicon extraction using such models has recently become very intense. Translation models have been improved in various ways and tools for statistical machine translation have become available in recent years. Estimation approaches using translation models have been incorporated into our word alignment approach described in section 5.1.3.

3 Compilation of parallel corpora

Parallel corpora are the essential data in this thesis. All our investigations depend on these collections of translation data. In this chapter, parallel corpora, which have been compiled while working on the thesis, are briefly described. Information about their origin, contents and annotation details are given below. Additional facts can be found in several publications and technical reports [Tie98, Tie99b, SFT+02, Ahr03, TN03].

3.1 Corpus encoding

Corpora can be stored in various formats. The most common form is to encode corpus data in structured text formats using markup languages such as the standard generalized markup language (SGML) or the extensible markup language (XML). In recent years, XML has succeeded SGML as the commonly used standard for the development of corpus encoding formats. The next three paragraphs introduce the XML-based formats, which have been used for the corpora described in the second section of this chapter.

3.1.1 PLUG XML

The PLUG project [Såg02] was started in co-operation with three Swedish universities, Gothenburg University, Linköping University and Uppsala University. The initial delivery from this project was defined as the compilation of a common project corpus, the PLUG corpus, with contributions from the three partners. The contributions comprise several bitexts collected at the three departments. They were delivered in three different formats. Gothenburg and Linköping used plain text formats for storing sentence aligned bitexts and Uppsala used the SGML based TEI[1] lite format for their sentence aligned data [TKS96]. For the PLUG corpus, a common format was defined for easy and efficient use during the project. The PLUG XML format defines a simple XML scheme for storing bilingual sentence aligned data in one XML file. The format is very similar to the translation memory exchange format TMX in its simplest form. In general, PLUG XML corpus files contain a short header and a collection of sub-corpora in the body. A sub-corpus may

[1] TEI is the Text Encoding Initiative (http://www.tei-c.org/).

include several documents containing sentence aligned segments (marked with "align" tags). Sub-corpora and documents include headers with descriptive information about their contents. The main contents is stored in very simple structures with only some basic markup. A short example is given in figure 3.1.

```
<?xml version="1.0"?>
<!DOCTYPE plug SYSTEM "dtd/plugXML.dtd">
<PLUG>
...
<align id="ensvfbell2" link="1-1">
  <seg lang="en">
    <s id="en2.1">Then hand luggage is opened.</s>
  </seg>
  <seg lang="sv">
    <s id="sv2.1">Sedan öppnas handbagaget.</s>
  </seg>
</align>
...
<align id="ensvfbell867" link="2-1">
  <seg lang="en">
     <s id="en867.1">You lean back with a cup of coffee to luxuriate
        in the Oriental conversation of an intelligent man.</s>
     <s id="en867.2">Immediately you are involved in a
        tormenting discussion.</s>
  </seg>
  <seg lang="sv">
     <s id="sv867.1">Man lutar sig bakåt med en kopp kaffe för att
        avnjuta en orientalisk människas intelligenta konversation,
        och omedelbart är man indragen i en plågsam diskussion.</s>
  </seg>
</align>
...
```

Figure 3.1: PLUG XML.

More information about PLUG XML and its document type definition (DTD) can be found in [Tie99b].

3.1.2 KOMA XML

The KOMA project [KOM01] is a follow-up project of PLUG and MATS[2] with the focus on corpus based machine translation. One of the main goals of the project is to integrate linguistic knowledge in the extraction of bilingual lexical data and the application of such data to machine translation. The PLUG XML format does not support additional markup at the lexical level

[2] http://stp.ling.uu.se/mats

and, therefore, a new XML based format was developed at the NLPLAB at Linköping University. KOMA XML is inspired by XCES formats, which are introduced below. Parallel data are stored in separate files. Source and target language documents are stored in different files using the "liu-mono" DTD[3], which is a simple general corpus encoding format very much like parts of the XCES document specifications. Corpus data are split into sentences, tokenized, and tagged. An example of a monolingual corpus file can be seen in figure 3.2.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE linCorpus SYSTEM "liu-mono.dtd">
<linCorpus>
  <linHeader></linHeader>
  <text>
    <body>
       ...
    <s id="s2">
      <w id="w33" relpos="1" base="then" func="meta" fa="&gt;5"
                  stag="EH" pos="ADV">Then</w>
      <w id="w34" relpos="2" base="hand" func="attr" fa="&gt;3"
                  stag="&gt;N" pos="N" msd="NOM-SG">hand</w>
      <w id="w35" relpos="3" base="luggage" func="subj" fa="&gt;4"
                  stag="NH" pos="N" msd="NOM-SG">luggage</w>
      <w id="w36" relpos="4" base="be" func="v-ch" fa="&gt;5"
                  stag="AUX" pos="V" msd="PRES-SG3">is</w>
      <w id="w37" relpos="5" base="open" func="main" fa="&gt;0"
                  stag="VP" pos="EN" msd="PASS">opened</w>
      <w id="w38" relpos="6" base="."
                  stag="INTERP" pos="INTERP" msd="Period">.</w>
    </s>
       ...
```

Figure 3.2: KOMA XML - a monolingual corpus file.

The alignments between source and target documents are stored in separate files using the "liu-align" DTD. Sentence links are collected as external links within link lists. Alignments in "liuAlign" files are hierarchical, i.e. word links are included as sub-structures of sentence links. Additionally, specific attributes for sentence links and word links, which match the needs and resources within the project, have been defined. Figure 3.3 shows a simple example of a bitext link file in KOMA XML.

More information about the KOMA XML formats can be found in [Ahr03] and in the KOMA XML DTDs.

[3]The prefixes "liu" (figure 3.2) and "lin" (figure 3.3) refer to Linköping University.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE liuAlign SYSTEM "liu-align.dtd">
<liuAlign fromDoc="svenprf.ces.src"
          toDoc="svenprf.ces.trg" version="1.0">
<liuHeader></liuHeader>
<linkList>
  ...
  <sentLink id="ensvfbell2" xtargets="s2 ; s2">
     <wordLink certainty="0.165" method="aut"
               xtargets="w36+w37;w29" lexPair="is opened;öppnas" />
     <wordLink certainty="0.268" method="aut"
               xtargets="w33;w28" lexPair="Then;Sedan" />
     <wordLink certainty="0.165" method="aut"
               xtargets="w34+w35;w30" lexPair="hand luggage;handbagaget" />
  </sentLink>
  ...
</linkList>
</liuAlign>
```

Figure 3.3: KOMA XML - an alignment file.

3.1.3 XCES

XCES is the XML version of the Corpus Encoding Standard that was developed at the Department of Computer Science at Vassar College, Poughkeepsie, New York and at Equipe Langue et Dialogue at LORIA/CNRS, Vandoeuvre-lès-Nancy, France [IPD00]. XCES provides specifications for several kinds of linguistically annotated corpora, among others also aligned multilingual data using the "cesAlign" document specifications. Similar to KOMA XML, bilingually aligned data can be stored in separate files: one for each monolingual corpus and one (or more) for the alignment between corpus files. The syntax is quite similar to the one in KOMA XML files. Links can be stored in so-called link groups ("linkGrp") within the alignment files. Link groups can be used for aligning any kind of unit, for instance sentences or words, but they cannot be hierarchical as in the "liu-align" DTD. An alignment file may look like the short example in figure 3.4.

XCES documents may use a variety of markup defined in separate DTDs. Several XCES DTDs may be linked together in order to include appropriate markup for specific tasks. More information about XCES can be found on its homepage (http://www.xml-ces.org/) and in the CES reference manuals [IPD00].

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign SYSTEM "xcesAlign.dtd">
<cesAlign>
  <linkList>
  <linkGrp fromDoc="svprf.ces" toDoc="enprf.ces">
    ...
    <link id="svenprf6" xtargets="s6 ; s6" />
    <link id="svenprf7" xtargets="s7 ; s7 s8" />
    ...
  </linkGrp>
  <linkGrp fromDoc="svpeu.ces" toDoc="enpeu.ces">
  ...
  </linkGrp>
</cesAlign>
```

Figure 3.4: XCES - an alignment file.

3.2 Corpora

The following sections briefly describe the five parallel corpora that have been compiled and used in the thesis.

3.2.1 The PLUG corpus

The quadri-lingual PLUG corpus contains approximately 2.2 million running words in 14 sub-corpora. It is the main resource for the alignment experiments presented in section 5.3. About half of the corpus consists of Swedish-English bitexts (in both directions). Other language pairs are Swedish-German and Swedish-Italian with about 500,000 words each. The corpus contains documents from three different genres, i.e. technical text, political text and literary text. The largest part is represented by technical documents comprising about 1.3 million words in total. Political and literary documents in the corpus contain about 400,000 words each. Swedish-English and Swedish-Italian bitexts are included in all three genres in at least one translation direction. There is no Swedish-German literary text in the corpus. Detailed information about the sub-corpora and their origin can be found in [Tie99b] and on the homepage of the PLUG project (http://stp.ling.uu.se/plug/).

3.2.2 The Scania 1998 corpus

The Scania 1998 corpus is an extension of the multilingual Scania corpus, which was initially compiled for eight European languages at Uppsala University. All documents have kindly been provided by the Swedish truck manufacturer Scania CV AB, Södertälje. The 1998 documents (mainly

in Swedish and English) were converted to Uppsala's SGML format (using TEI lite) and translated documents have been sentence aligned using Gale&Church's length based alignment program [Chu93]. Scania 1998 contains over 2 million words of Swedish and English (in mainly parallel documents). The Scania corpus has mainly been used for lexicographic work and for compiling the Scania lexicon, which is used in the controlled language checker *ScaniaChecker* [AS00] and in the machine translation system MATS [SFT+02]. More information about the corpus can be found in [Tie98] and on the homepage of the Scania project (http://stp.ling.uu.se/scania).

3.2.3 The MATS corpus

The MATS corpus is a tri-lingual corpus of sentence aligned technical documents, which has been compiled for the machine translation project MATS [SAJ+99]. The corpus contains Swedish-English and Swedish-German bitexts from Scania with about 100,000 words for each language pair. The corpus is encoded in sentence aligned TEI lite format and in PLUG XML format. The Swedish part of the corpus has been partially tagged and analyzed syntactically [LW01]. The corpus has been split into a training corpus and a validation corpus and is mainly used for the development of a domain-specific machine translation system using the MATS MT platform [SFT+02].

3.2.4 The KOMA corpus

The KOMA corpus is a compilation of Swedish-English bitexts from the two partners within the corpus-based machine translation project KOMA: the NLPLAB at Linköping University and the Department of Linguistics at Uppsala University. The corpus includes technical manuals from the PLUG corpus, the MATS corpus, and two additional technical manuals. It has been converted to the KOMA XML format and tagged using Connexor's Functional Dependency Grammar (FDG) tagger (http://www.connexor.com/).

The corpus contains six sub-corpora with a total of 1.6 million words. It is used for automatic and interactive word alignment [AMA02, Tie03] and for the development of machine translation [SFT+02, Hol03]. More information about the corpus can be found in [Ahr03] and on the project homepage (http://www.ida.liu.se/~nlplab/koma/).

3.2.5 The OPUS corpus

OPUS [TN03] is a growing multilingual corpus of translated open source documents available on the Internet. In the current version (v 0.2), the corpus includes about 30 million words in 60 languages which have been collected from

three sources: OpenOffice.org[4] documentation (http://www.openoffice.org), PHP[5] manuals (http://www.php.net/download-docs.php) and KDE[6] manuals including KDE system messages (http://i18n.kde.org). The OpenOffice.org sub-corpus (OO) contains about 2.6 million words in six languages. The corpus is completely parallel, i.e. all English source documents have been completely translated into five languages. The KDE manual sub-corpus (KDEdoc) includes 24 languages with about 3.8 million words in total. The translation initiative at KDE is an on-going project. Hence, documents are, so far, only partly translated for many languages. New languages are added constantly. KDE system messages have been compiled into a separate sub-corpus (KDE) containing about 20 million words in 60 languages. Even this translation initiative is on-going and many languages have been translated only in parts. The sub-corpus of PHP manuals (PHP) is derived from the HTML version of the on-line documentation of the scripting language PHP. PHP contains about 3.5 million words in total in 21 languages. The sub-corpus is rather noisy as non-translated parts are kept in the original English version in the HTML documents. However, most of this noise in the PHP corpus has been removed automatically.

The main motivation for compiling OPUS is to provide an open source parallel corpus that uses standard encoding formats and as much additional linguistic information as possible. All corpus files have been encoded in Unicode UTF8 and sentence aligned for all possible language pairs (e.g. 1830 language pairs for KDE). Sentence alignments are stored in XCES format as described above. Corpus files are stored in XML using the original markup from the source documents with added linguistic markup. Additional markup includes sentence boundaries (for all documents as it is needed for sentence alignment), word boundaries (for all languages except Asian languages such as Chinese for which no tokenizer was available), part-of-speech tags (for English, French, German, Italian, Swedish in parts of the corpus) and shallow syntactic structures (for English in parts of the corpus). We are grateful for the tools that have been provided by external researchers for adding linguistic markup to the corpus in its current version [Bal02, Bra00, Sch94, MKY$^+$00, Meg02]. More information will be added gradually when tools become accessible to us. Figure 3.5 shows an example of linguistically enriched corpus data from the OPUS corpus.

The original markup provides structural information such as paragraph boundaries, headers, lists and tables. Maintaining the original markup and

[4]OpenOffice.org is an open source office suite.

[5]PHP:Hypertext Preprocessor (PHP) is a widely-used general-purpose scripting language which is available as open source.

[6]The K Desktop Environment (KDE) is free graphical desktop environment for UNIX workstations.

```
<ul class="L2">
  <li class="">
    <p class="P4" id="8">
      <s id="s8.1">
        <chunk id="c8.1-1" type="NP">
          <w grok="NNP" tree="RB" lem="over" tnt="IN">OVER</w>
        </chunk>
      </s>
    </p>
    <p class="P5" id="9">
      <s id="s9.1">
        <chunk id="c9.1-1" type="NP">
          <w grok="NNP" tree="VBP" lem="overwrite" tnt="NNP">
             Overwrite</w>
          <w grok="NN" tree="NN" lem="mode" tnt="NN">mode</w>
        </chunk>
        <chunk id="c9.1-2" type="VP">
          <w grok="VBZ" tree="VBZ" lem="be" tnt="VBZ">is</w>
          <w grok="VBN" tree="VBN" lem="enable" tnt="VBN">
             enabled</w>
        </chunk>
        <w grok="." tree="SENT" lem="." tnt=".">.</w>
      </s>
```

Figure 3.5: Linguistic markup in OPUS.

the original document structure makes it possible to go back to the original source, makes it easy to produce sub-sets of the corpus, and also increases the performance of the automatic sentence alignment by reducing follow-up errors. Furthermore, using several tools for similar tasks such as part-of-speech tagging makes it possible to identify "weak" points in the automatic annotation (see, for example, the three part-of-speech tags (grok, tree, tnt) for "OVER" and "Overwrite" in figure 3.5).

The entire corpus is freely available from the homepage of the project (http://logos.uio.no/opus/). It can be downloaded in its native XML source format or compiled as sentence aligned HTML-documents. Additional sub-corpora will be added continuously. More information can be found in [TN03] and on the homepage of the project.

3.3 Summary

In this chapter, our corpus data and their encoding formats have been presented. Five parallel corpora have been compiled within the research of this thesis, in co-operation with our project partners. However, it is mainly the PLUG corpus that is used in our alignment investigations (chapter 5). This

corpus includes small bitexts from different genres that makes it perfectly suitable for comparative experiments. Furthermore, many tools have been tailored towards this corpus within the PLUG project. Three bitexts from the PLUG corpus are used in our word alignment experiments presented in chapter 5. Most of the other corpora have been processed in a similar way but results have not been investigated and evaluated in the same manner. The MATS corpus is used in a study on lexicography (section 6.1.1), in experiments with a prototype for interactive translation (section 6.2.1), and in the development of a machine translation system (section 6.2.2). The Scania1998 corpus is used in another study on lexicography (section 6.1.2) and also for the augmentation of the Scania lexicon used for our research on machine translation (section 6.2.2).

4 Development of tools

Tools are required for processing parallel data. In this chapter, I will present the software that has been developed and used in the thesis. The chapter includes a brief description of the Uplug toolbox, a note on recent modifications of this implementation, and a description of corpus interfaces and other tools. The last section contains information about a lexical database for storing refined bilingual lexicons that have been extracted from parallel corpora.

4.1 Uplug

Uplug is a general corpus processing tool available in two versions. For simplification, we will call the original version *Uplug I*, which refers to the system that has been developed in the PLUG project [Tie02b]. The new version of the toolbox will be called *Uplug II*. Differences between them will be explained below.

Uplug I

The Uplug toolbox was developed as a general platform for corpus processing within the PLUG project. The task of this toolbox is to provide a common interface for the combination of several modules with different data needs and formats. Consequently, Uplug was designed to be a modular extensible toolbox with three components:

UplugSystem: a launcher for combining modules in sequential processes (pipeline architecture)

UplugIO: an extensible input/output interface for record-based data with flat or shallow structures

UplugGUI : a graphical interface for parameter adjustments and data inspection

An overview of the system is presented in figure 4.1.

Uplug modules are organized in sequences with access to data via the UplugIO interface. Data streams are collections of sequential data, which can be read record-wise from a file, database, or any other source that is supported

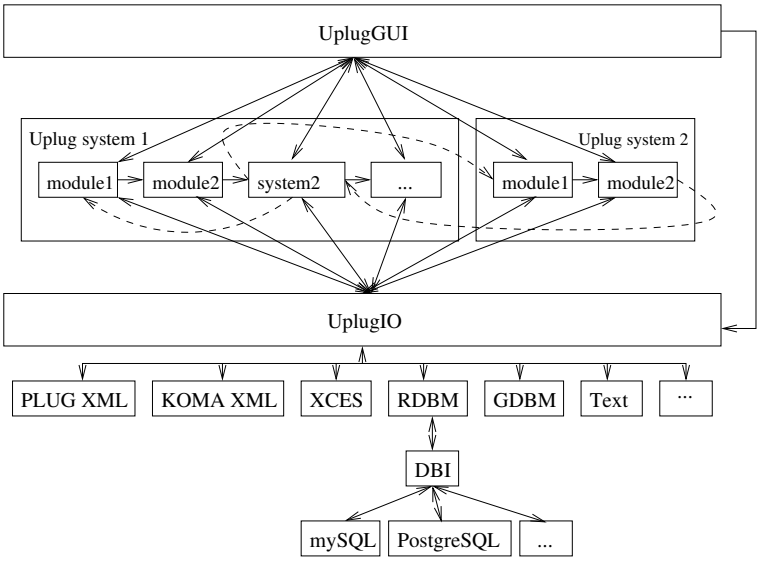


Figure 4.1: The design of the Uplug toolbox.

by UplugIO. Results of each module may be pushed into the next module in the sequence until the final module is reached. It is also possible to define multiple iterations for module sub-sequences, which are specified as "jumps" back to previous modules. Data are transferred in the same way as in normal module sequences. Uplug modules may be other Uplug systems as well. Hereby, hierarchical system structures can be defined, i.e. sequences of modules that call other sequences of modules. System 2 in figure 4.1 is an example of a sub-system that is embedded in system 1.

The idea behind the UplugIO is to provide a transparent interface to data from different sources so that individual modules are not concerned with the actual storage format of their input and output. Data records can be read from several sources (data streams) and modules receive them in a standard format independent of the external format. UplugIO also supports collections of data sources even with different external storage formats such that several sources can be combined into one stream with a common data structure.

Uplug systems, modules and their connections to the UplugIO interface and the graphical interface are defined in configuration files. The syntax of configuration files in the original implementation is described in detail in [Tie02b]. This implementation (Uplug I) has been used for the PLUG Word Aligner (PWA), a package combining two aligners, i.e. the Linköping Word Aligner (LWA) [AMA98] and the Uppsala Word Aligner (UWA). The package includes also a few other modules, for instance, a module

for extracting collocations and a module for evaluating word alignment results using pre-defined gold standards. PWA is freely available for research purposes. More information can be found on the PWA homepage (http://stp.ling.uu.se/plug/pwa/).

Uplug II

Uplug I was implemented to handle data with shallow structures. It mainly supports flat, record-based data structures. In UWA, corpus data are stored internally in a Tipster like format [Gri98], i.e. the original corpus is saved as a plain text file and annotations are stored externally using pointers back to the document. The Tipster architecture is very flexible as many layers of annotations can be added without changing the original text document. However, there are disadvantages, for example, that a text document may never be changed after annotation. Any change in the text document, intended or accidental, may damage the complete annotation that refers to parts following the point of modification. Furthermore, Tipster annotations are not readable without appropriate tools, which present text and annotation together in a adequate way as it is done, for instance, in the GATE toolbox [GRCH96].

The main motivations for a new implementation of Uplug were the following: First, the new system ought to be able to handle deeply structured data. It is necessary to move easily within such data structures in order to explore their contents and internal relations. Secondly, the system should be applicable to many language pairs making it necessary to work with Unicode. The following changes are implemented in the Uplug II system:

- Uplug II uses the XML Document Object Model (DOM) as a default for handling structured data. DOM data can be used for XML documents and for other data. Uplug II applies a standard implementation of the DOM interface using Perl.
- DOM data are internally stored in Unicode. Consequently, Uplug II works with Unicode as its default. However, many other character encoding standards are supported, for example, the common 8-bit character sets of the ISO-8859 standards.
- Uplug II is implemented in an object-oriented style. Object-oriented programming is a natural choice for integrating input/output interfaces for different formats. Re-usability of code is in general a native feature of the object-oriented programming paradigm.
- The syntax of the configuration files has changed. Uplug II uses Perl hashes dumped into plain text files for storing configuration parameters. This makes configuration structures more flexible and easier to validate using Perl itself for parsing the files.

The design of Uplug II is the same as in Uplug I but the implementation is new. Both systems are implemented in Perl, which makes them flexible but

slow. The integration of structured data and the complex DOM interface slow the system down even more. However, Uplug II is a very useful platform for experiments on corpus data, parallel or non-parallel, structured or plain text. Uplug II is currently in the beta stage and is not available in the same form as version I. The main application is the Clue Aligner, which is introduced in section 5.1.3, but it is also used as a general corpus tool for tasks like sentence splitting, tokenization, format conversions, and as a wrapper for tools such as taggers, shallow parsers, and the sentence aligner.

4.2 Interfaces and other corpus tools

Additional interfaces and tools besides Uplug have been developed for the work with parallel corpora. Several web-based search interfaces have been implemented for the inspection of corpus data. A bilingual concordancer has been developed that can be used to query bilingually aligned parallel corpora. It uses the input/output interface of the Uplug toolbox and the XML string search tool *sggrep* which is part of the LT XML tool collection from Edinburgh [BMT+00]. All bitexts from the PLUG corpus have been integrated in the concordance tool. Furthermore, Swedish-English, Swedish-German, and Swedish-Italian corpora from the Scania corpora can be searched via the same interface. A monolingual concordance tool has also been implemented for the Swedish part of the Scania corpus and is also available with restricted access on the web. A sample query is shown in appendix A.1. Multilingual search interfaces have been developed for parts of the OPUS corpus. These web-interfaces are freely accessible from the OPUS homepage and use the Corpus Workbench (CWB), which has been developed at the Institute for Natural Language Processing (IMS) in Stuttgart [Chr94]. The OPUS query tools allow the search of several languages in parallel using the powerful corpus query language (CQP) which is implemented in the CWB. Another advantage of the OPUS tools is the possibility to search for annotations such as part-of-speech tags. A sample query is shown in appendix A.2.

Other tools that have been developed include link databases for the storage of word alignment results and an experimental web-based interface to Uplug modules (UplugWeb) presented at NODALIDA 2001 [Tie01c]. UplugWeb makes it possible to run Uplug modules with small data sets via web interfaces, which makes it possible to run the system remotely from external hosts. UplugWeb includes basic modules and querying tools for searching corpus data and alignment results. A web-based interface to the link database is integrated in the system as well. The link database applies a relational database for fast and easy access to collected word links.

4.3 MatsLex - a lexical database

The multilingual lexical database MatsLex has been developed within the MATS project [MAT00] and the on-going KOMA project on machine translation. One of the main goals of MATS and KOMA is to scale-up an existing transfer-based machine translation system by integrating lexical data extracted from parallel corpora. MatsLex was designed to provide a central database for all the lexical information employed by the machine translation system. The structure of the database allows the storage of many kinds of linguistic information about lexical items in different languages and to link lexemes from two languages together. MatsLex applies a relational database management system (RDBMS) and includes several interrelated tables for each language concerned. New languages can easily be added using identical structures. More details about the database structure can be found in [Tie02a].

MatsLex is used to manage the lexical data in a central multi-user environment. The database can be processed using web-based interfaces and command-line tools. Lexical items in the database are basically stored in the form of related lexemes, lemmas and inflectional patterns. Several syntactic and semantic features can be added to each item in the database. Word forms are not explicitly stored in MatsLex. They are generated from technical stems and substitution patterns specified for each inflectional pattern. The main advantage of this technique is to simplify the extension of the lexicon. New words can be added to the database using their base forms and linking them to a corresponding inflectional pattern. In this way, all possible word forms according to the pattern are implicitly included for each base form in the lexicon. This enforces consistency throughout lexicon entries with respect to inflectional patterns and their associated features. A disadvantage of this technique is the large amount of inflectional patterns that have to be known by the database user when adding new words to the database. Tools may help to find appropriate patterns for new database entries. A prototype of a *pattern finder* has been implemented in an earlier study [SAS97].

Another distinctive feature of the database is the possibility of using regular expressions that can be stored in order to represent classes of similar tokens such as dates, time-expressions and numbers. Regular expressions are used in the same way as ordinary items in the lexical database. They are linked to substitution patterns, which translate matching items into correct correspondences in the target language. More details are presented in [Tie02a].

The connection of the database to the machine translation system is implemented via compiled run-time lexicons produced from the MatsLex database. Run-time lexicons are read-only and contain all the data necessary for the translation system. The main reasons for compiling run-time lexicons is to separate the lexical components of the translation system from the dynamic

contents in the database and to provide optimized data structures for efficient access by the system. MatsLex is used in a multi-user and multi-tasking environment. It can be accessed and updated simultaneously by several users. Direct access from the machine translation system to the database would be possible but the dynamic nature of the database may cause confusion in the development and usage of the translation system. Run-time lexicons are fixed "snap-shots" of the database that represent a certain stage of the lexicon. In this way, authorized versions of the lexicon can be extracted from the database and used consistently. Multiple run-time lexicons can be kept to represent different stages of the database. This makes it easy to compare results with earlier versions and to backtrack in case of errors. Furthermore, the components of the translation system require different sub-sets of data. Searching the database for specific data and converting them to the required format (e.g. creating word forms from technical stems and inflectional patterns) can be time-consuming and inefficient. Run-time lexicons can be compiled whenever a stable lexicon has been collected and their contents can be accessed efficiently. More details about the connection between MatsLex and the machine translation system can be found in [Tie02a].

4.4 Summary

Two kinds of tools have been presented in this chapter: general corpus processing tools and tools for handling lexical databases. The Uplug corpus tool originates from the PLUG project as one of Uppsala's contributions. Its name refers to the modular design of this toolbox. Uplug is meant to serve as a general platform for "plugging" modules together as one pleases ("you plug"). Uplug has been implemented in two different versions. The first one is freely available for research purposes and includes mainly the two word aligners (UWA and LWA) developed within the PLUG project. The second version generally uses the same architecture but is tailored towards the Clue Aligner presented in section 5.1.3. MatsLex originates from the MATS project and implements tools for handling lexical data in a relational database. MatsLex includes several interfaces for the work with the database and its connection to a machine translation system. MatsLex is included in the web-based machine translation platform MATS, which is under development at the Department of Linguistics at Uppsala University.

5 Word alignment strategies and experiments

Word alignment is the basic task in extracting lexical data from parallel corpora. In this chapter, we will concentrate on techniques for automatic word alignment that have been investigated, developed and implemented in the thesis.

The task of word alignment was introduced in section 2.3. We will emphasize the combination of multiple resources for achieving alignment accuracy. In the first section, word alignment techniques are described. It includes an overview of alignment resources and association measures, and a presentation of two word alignment systems that have been developed in the thesis (the Uppsala Word Aligner (UWA) and the Clue Aligner). The second section of this chapter describes metrics for word alignment evaluation including a proposal of a new refined measure for the evaluation of partially correct alignment results. Finally, the last section, before summarizing the chapter, presents experimental results that have been achieved with the word alignment techniques as described in the sections before.

5.1 Alignment strategies

In this section we will discuss in detail, word alignment techniques used in our investigations. The first part summarizes resources and measures. The second part presents the principles of the Uppsala Word Aligner (UWA) and its "greedy" alignment approach. The third part contains a detailed description of the clue alignment approach including definitions of the terminology and explanatory examples.

5.1.1 Alignment resources

The main resource in word alignment is the bitext itself. Correspondences between words and phrases in the two languages can be derived directly from the corpus using statistics of word distributions and measures of similarity. Another group of resources for word alignment are external knowledge sources which can be divided into machine-readable data collections (such as bilingual dictionaries) and expert knowledge (such as association heuristics between word positions or part-of-speech tags).

Statistical resources

Statistical approaches to word alignment are usually built upon the co-occurrence of words and their translations in bitext segments. Statistical techniques as introduced in section 2.3.2 capture the co-occurrence property by estimating alignment parameters that describe translation relations according to a translation model. The estimated parameters maximize the likelihood of the data given the model, i.e. recurrent word pairs with similar contextual features obtain high probabilities when maximizing the likelihood function. More details about statistical estimation approaches can be found in chapter 2. Association approaches to word alignment use measures of correspondence for the purpose of identifying translation relations. These association measures are often derived from statistics using co-occurrence frequencies as their main parameter. Section 2.3.1 gives some background to such measures that have been applied in alignment tasks. In our approaches, we apply three co-occurrence measures, the Dice coefficient, t-scores, and point-wise mutual information. More information about these measures can be found in section 2.3.1.

String matching techniques

String similarity measures are used to find cognates in bitexts. Two common measures have been presented in section 2.3.1, the Dice coefficient and the longest common sub-sequence ratio (LCSR). In our alignment approaches, we apply the LCSR score.

LCSR measures string similarity based on common characters. However, the spelling of cognates often differs systematically between languages even if they are close to each other phonologically. Loan words that enter a language are usually altered in order to match language specific structures and rules. In many languages certain modifications are consistently applied when borrowing words from other languages. For example, the consonant 'c' in English usually corresponds to the letter 'k' in Swedish (and German) and vice versa when representing the phoneme /k/. Other examples of systematic spelling changes are the French word "chauffeur", which became "chaffför" in Swedish, and the English word "mail", which is often spelled "mejl" in Swedish. Sometimes language specific characters are simply replaced by similar characters for being consistent with the alphabet of the borrowing language, for instance, "smorgasbord" in English, which is borrowed from Swedish ("smörgåsbord").

A cognate recognizer should be aware of systematic spelling differences in order to identify a large number of possible cognates. Weighted string matching algorithms are discussed in [Tie99a] which can capture similarities even between non-identical characters. The same algorithm as for standard LCSR is used together with a character mapping function. This function defines weights for matching pairs of characters, even non-identical ones.

In [Tie99a], three approaches the automatic construction of weighted string similarity measures are proposed. Weighted character mapping functions are learned automatically from a collection of cognate candidates. In the first approach pairs of vowels and pairs of consonants at similar positions are mapped to each other and co-occurrence scores are calculated for each character pair using the Dice coefficient. These scores represent the weights in the character mapping function. Similarly, in the second approach, pairs of vowel sequences and pairs of consonant sequences are mapped to each other and Dice scores are calculated in order to fill the character mapping function. Note that the unit to be matched is now a sequence of characters, which may consist of more than one character. This is also the case for the third approach, which yields the best result in the presented experiments. It is based on the co-occurrence of non-matching parts at similar positions in previously extracted cognate candidates. The following example illustrates non-matching parts from a Swedish-English cognate candidate:

Swedish	*k*	r	i	t	i	*sk*	a	
English	*c*	r	i	t	i	*c*	a	*l*

Non-matching parts in this example are the pairs ('k' → 'c'), ('sk' → 'c'), and ('' → 'l'). In [Tie99a], it has been demonstrated experimentally for Swedish and English, that a weighted string similarity measure with a mapping function that has been learned using the third approach improves recall and even precision for cognate extraction tasks.

Other resources

In section 2.3.1, external resources for word alignment have been discussed. Several resources have been applied in our word alignment approaches. Machine-readable bilingual dictionaries (MRBD) and alignment heuristics such as position weights have been used in the Uppsala Word Aligner which is presented in the following section. Pre-defined relations between morphosyntactic features are applied in the Clue Aligner which is presented in section 5.1.3.

5.1.2 Greedy word alignment - step by step

UWA implements a "greedy" word alignment approach based on association measures and alignment heuristics. The basic idea in this approach is to combine different resources in an iterative alignment procedure. The principles of this approach are inspired by Martin Kay's quote on machine translation [Kay97, p. 13]:

> ”The keynote will be modesty. At each stage, we will do only what we know we can do reliably. Little steps for little feet!”

Word alignment is a non-trivial task and also walks on “little feet”, which leads to the principles of “baby-steps” for word alignment as formulated in [Tie99c, p. 218]:

1. Prepare carefully before taking the first step.
2. Use all available tools that can help.
3. Check alternatives before taking the next step.
4. Take safe steps first; try risky steps later.
5. Remove everything that is in the way.
6. Improve walking by learning from previous steps.
7. Reach the goal with many steps rather than one big one.
8. Continue trying until you cannot get any closer.

Using these principles and the modular design of the Uplug system, the Uppsala Word Aligner (UWA) was designed as sketched in figure 5.1.

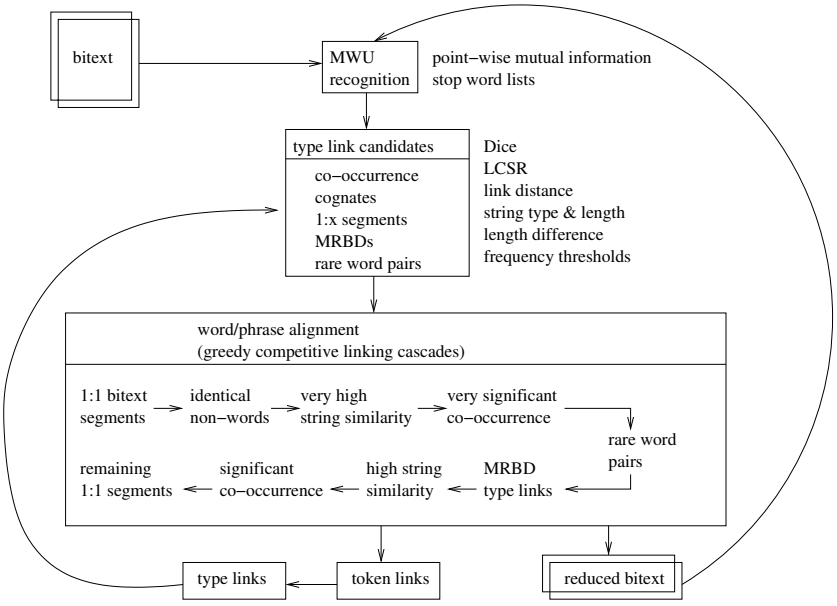


Figure 5.1: The Uppsala Word Aligner.

The aligner starts with a pre-processing step (principle 1), which includes the recognition and extraction of MWUs for both languages in the bitext. Collocations are identified using co-occurrence statistics (point-wise mutual information) and language-specific classified stop word lists for phrase boundary detection. In the second step, alignment candidates by means of

type links are collected from any appropriate source (principle 2). Candidates are ranked according to their reliability (principle 3), i.e. their association score. Alignment candidates include cognates (found by means LCSR scores), associated pairs (found by means of Dice scores, point-wise mutual information or t-scores), single-word bitext segments[1], pairs of rare words (which occur significantly less frequently than all other words in the bitext segment), and machine-readable bilingual dictionaries (MRBD). "Risky" candidates should be avoided (principle 4); several heuristics help to keep the collection of link candidates as clean as possible: Link distance thresholds (assuming that corresponding words do not occur arbitrarily far apart from each other), string length thresholds (short words are unreliable for string similarity measures), constraints on length difference ratios (assuming that related words do not exceed a certain length difference), and frequency thresholds (sparse data produce unreliable statistics). Word alignment includes several steps (principle 7) in the form of linking cascades. The process is implemented as a greedy, best-first (principle 4) search in the manner of competitive linking. Competitive linking implies that aligned words cannot be aligned again which allows the removal of linked words (principle 5). Unlike Melamed's approach [Mel96a], we do not restrict the greedy search to one-to-one alignments but include also (possibly overlapping) multi-word-units (MWUs) in the word alignment competition.

The word alignment cascades are not the end of the process. Type links are extracted from the aligned bitext and return to the alignment process as an additional resource (principle 6). Due to the competitive linking process the size of the corpus is reduced by the number of aligned words. The reduced corpus enters another alignment iteration starting, again, with a "careful" preparation of the first step in the process, i.e. the collocation extraction. The MWU recognizer finds adjacent collocations only. However, having aligned words removed from the bitext it is possible to find long-distance collocations, assuming that interrupting items have been removed completely from the corpus in the previous iteration. This motivates the repeated use of the MWU recognizer. Furthermore, the contents of the corpus has been changed significantly with respect to word occurrence and co-occurrence of word pairs. Statistically significant associations between words (and new MWUs) may arise in the reduced bitext, which could not been identified previously in the corpus. Finally, the reduced bitext may contain additional single-word bitext segments that can be directly aligned. The iterative process can be continued as long as new links are found (principle 8).

[1] Single-word bitext segments denote bitext segments with only one word on either the source or target language segment. Sentence alignment may produce many of them by aligning, for instance, table cells.

5.1.3 Combining clues for word alignment

Word alignment using a combination of association clues has been introduced in [Tie03]. It is inspired by previous work (as described in section 5.1.2) but applies refined techniques for the combination of different resources. The main idea about the clue alignment approach is to incorporate several knowledge sources into the word alignment process in a systematic way. Empirical data shall be complemented with linguistic information, statistical parameters shall be combined with alignment heuristics.

Most word alignment models apply knowledge-poor approaches, i.e. alignment based on plain texts disregarding all additional information that can be gathered for the languages involved. Simple tools such as stemming functions and language-specific stop word lists are often added to improve the performance of simple association approaches. Recently, initial studies on the impact of linguistic information, such as part-of-speech, on word alignment have been carried out as described in chapter 2. The clue alignment approach represents a flexible framework for such a combination of language-specific, heuristic, and statistical resources in word alignment.

Word alignment clues

For many languages, reliable tools are available for enriching corpus data with linguistic information. Examples of such tools are part-of-speech taggers, lemmatizers, named-entity recognizers, and (shallow) syntactic parsers. Other kinds of language-specific information may be found in dictionaries including idiomatic expressions, multi-word terms, word meanings, word translations etc.

It may be assumed that such data and tools have a positive impact on the identification of translation relations between lexical units in bitexts. For an illustration of this intuition, see for instance the simple example in figure 5.2 where translation correspondences are expressed between automatically tagged chunks rather than between words.

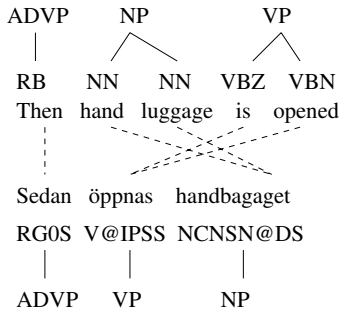


Figure 5.2: Linguistically enriched bitext segments.

Linguistic information and contextual features may be regarded as *clues* to relations between words. In particular, these clues may indicate translational correspondences between words in bitexts. Contextual clues are frequently used by lexicographers for the identification of word senses. Similarly, the clue alignment approach tries to utilize information about the context for the identification of translation equivalents. Many types of features can be explored: Spelling similarities may indicate cognates. Information about word classes may support linking decisions. Word position can be another key to alignment preferences. Morphosyntactic features and information about syntactic functions may help to find structural similarities and translational correspondences between lexical items within these structures. Some of these *alignment clues* may be very specific (e.g. default translations taken from a dictionary) and others may be very general (e.g. pairs of chunk labels such as 'NP-NP' in figure 5.2). Depending on their impact, clues may be of different importance for the identification of translational correspondence.

The following definitions are used in the clue alignment model:

Word alignment clue: A word alignment clue $C_i(s,t)$ is a probability $P(a_i|s,t)$ which indicates an association a_i between two lexical items s and t in parallel texts. The association is used as a binary function, i.e. the probability distribution includes only the two values $P(a_i|s,t)$ (representing the probability of s and t being associated) and $P(\neg a_i|s,t) = 1 - P(a_i|s,t)$ (representing the probability of s and t not being associated).

Lexical item: A lexical item x is a set of words[2] with associated features f_x attached to it. Features may include any information attached to x or to the context of x (word position may also be a feature).

Declarative clue: Clues which have been pre-defined are declarative clues, i.e. declarative clues are independent of the training data. Typical examples of declarative clues are pre-defined associations between word pairs from related word classes (represented by their part-of-speech tag) or translations derived from machine-readable dictionaries.

Estimated clue: Clues which have been derived from training data are called estimated clues. They can be derived from association measures: $C_i(s,t) = w_i A_i(s,t)$ with w_i being a factor for weighting and normalizing the association score and $A_i(s,t)$ being a measure such as the Dice coefficient $A_{Dice}(s,t) = w_{Dice} Dice(s,t)$ or the longest common sub-sequence ratio $A_{LCSR}(s,t) = w_{LCSR} LCSR(s,t)$. Other estimated clues can be learned from aligned training data: $C_j(s,t) = w_j A_j(f_s, f_t)$ where f_s and f_t are features of source and target language items, and w_j is a weight.

[2]We use sets in order to include MWUs in the definition. Word order is not explicitly defined but may be used as a feature.

Clue resources : A source of alignment clues or an association metric for clue estimation is called a clue resource.

Clue patterns : Patterns that match lexical or contextual features are called clue patterns. They are used for learning clues from aligned training data.

Total clue : Clues can be combined and the combination of all available clues is called the total clue: $C_{all}(s,t) = P(a_{all}|s,t) = P(a_1 \cup a_2 \cup .. \cup a_x|s,t)$ Clues are not mutually exclusive. Several association types can be found together, e.g., an association based on co-occurrence can be found together with an association based on string similarity. The total clue combines these associations in order to strengthen the evidence of certain translation relations indicated by different clues. The union of two clues is defined as follows: $P(a_1 \cup a_2|s,t) = P(a_1|s,t) + P(a_2|s,t) - P(a_1 \cap a_2|s,t)$. For simplicity, we assume that clues are independent of each other: $P(a_1 \cap a_2|s,t) = P(a_1|s,t)P(a_2|s,t)$.

Clue value distribution: A clue indicates associations between all its member token pairs. The distribution of clue values to member token pairs can formally be expressed as $C_{i,s^N,t^M}(s_n,t_m) = C_i(s^N,t^M)$ for the source item $s^N = s_1 s_2 .. s_N$ ($n \in \{1..N\}$) and the target item $t^M = t_1 t_2 .. t_M$ ($m \in \{1..M\}$).

The definition of word alignment clues is very general and allows a large variety of clue types. Clues are (as in real life) not always helpful. Bad clues can be misleading. This has to be considered when choosing clue resources and designing clue patterns. The impact of certain clues in combination with others is another issue, which becomes very important when setting weights for individual clue patterns and resources. The optimal solution for the compilation of alignment clues would be to find a way of automatic approval of possible clues for the task of word alignment including an automatic setting of optimal weights for "good" clues. This could be achieved using representative test data and an optimization procedure. However, this is very time consuming due to the large number of possible options and parameters. An intermediate solution is to adjust the clue settings experimentally according to the alignment results using intuitively reasonable settings and clue patterns.

The clue matrix

The distributional property of clues as defined above enables the combination of clues on a word-to-word level according to the clue combination rule. In this way, a value for the total clue can be calculated for each pair of words for a given bitext segment using all available clues, even for overlapping items. One-to-one relations between words of a bitext segment can be represented in a two-dimensional space, which we will call the *clue matrix* as it is filled with clue values. The following example is constructed from a bitext segment

taken from a translation of Saul Bellow's "To Jerusalem and back: a personal account" [Bel76] to Swedish [Bel77] (henceforth the *Bellow corpus*):

```
Then hand luggage is opened.
Sedan öppnas handbagaget.
```

Let us assume clues derived from co-occurrence measures (Dice) and string similarity measures (LCSR) in table 5.1.

co-occurrence (Dice)			string similarity (LCSR)		
then	sedan	0.38	hand luggage	handbagaget	0.67
is opened	öppnas	0.30	opened	öppnas	0.33
is opened	sedan öppnas	0.30	then	sedan	0.40
opened	öppnas	0.65	hand	sedan	0.40
luggage	handbagage	0.45	hand	handbagaget	0.36

Table 5.1: Word alignment clues.

Using the clues from table 5.1 we can fill each cell of the following clue matrix (figure 5.3) with corresponding values (without weighting scores).

	sedan	öppnas	handbagaget
then	63	0	0
hand	40	0	79
baggage	0	0	82
is	30	51	0
opened	30	83	0

Figure 5.3: A clue matrix (all values in %).

The example in figure 5.3 illustrates how overlapping clues may strengthen the reliability of word links. Most of the clues in the example do not strongly suggest links on their own (especially clues below 0.5). However, in combination they create a clearer picture about the alignment that should be preferred.

In the following section, word alignment strategies using association clues collected in clue matrices are presented and discussed.

Clue alignment

A clue matrix summarizes information from various sources that can be used for the identification of translation relations. However, there is no obvious way to utilize this information for word alignment as we explicitly include multi-word units in our approach. The clue matrix in figure 5.4 has been obtained for

a bitext segment from the Bellow corpus using a set of weighted declarative and estimated clues.

	ingen	visar	särskilt	mycket	tålamod
no	29	0	0	1	9
one	16	2	1	1	13
is	1	13	1	2	0
very	0	2	18	17	1
patient	2	1	4	12	6

Figure 5.4: Another clue matrix (all values in %).

There are many ways of "clustering" words together and there is no obvious maximization procedure for finding the alignment optimum. The alignment procedure depends very much on the definition of an optimal alignment. The best alignment for our example would probably be the set of the following links:

$$\text{links} = \left\{ \begin{array}{ll} \text{no one} & \text{ingen} \\ \text{is patient} & \text{visar tålamod} \\ \text{very} & \text{särskilt mycket} \end{array} \right\}$$

A typical procedure for automatic word alignment is to start with one-to-one word links. Links that have a common source or target language words are called *overlapping links*. Sets of overlapping links, which do not overlap with any other link outside the set, are called *link clusters*. Aligning words one by one often produces overlaps and in this way implicitly creates aligned multi-word-units as part of link clusters. A general word-to-word alignment L for a given bitext segment with N source language words ($s_1 s_2 ... s_N$) and M target language words ($t_1 t_2 ... t_M$) can be formally described as a set of links L_x

$$L = \{L_1, L_2, ..., L_x\} \text{ with } L_x = [s_{x_1}, t_{x_2}], x_1 \in \{1..N\}, x_2 \in \{1..M\}$$

This general definition allows varying numbers of links ($0 \leq x \leq N * M$) within possible alignments L. There is no obvious function that can be maximized for finding the optimal alignment. The notion of an alignment optimum can be defined in several ways.

One such word-to-word alignment approach is to assume a *directional word alignment model* similar to the models in statistical machine translation (see section 2.3.2). The directional alignment model assumes that there is at most one link for each source language word. Using alignment clues, this can be expressed as the following optimization problem: $\hat{L}^D =$

$argmax_{L^D} \prod_{n=1}^{N} C_{all}(L_n^D)$ where $L^D = \{L_1^D, L_2^D, .., L_N^D\}$ is a set of links $L_n^D = \left[s_n, t_{a_n^D}\right]$ with $a_n^D \in \{1..M\}$ and $C_{all}(L_n^D)$ is the "total clue value" for the linked items s_n and $t_{a_n^D}$. In other words, word alignment is the search for the best link for each source language word. Directional models do not allow multiple links from one item to several target items. However, target items can be linked to multiple source language words as several source language words can be aligned to the same target language word. The direction of alignment can easily be reversed, which leads to the *inverse directional alignment*: $\hat{L}^I = argmax_{L^I} \prod_{m=1}^{M} C_{all}(L_m^I)$ with links $L_m^I = \left[s_{a_m^I}, t_m\right]$ and $a_m^I \in \{1..N\}$. In the inverse directional alignment, source language words can be linked to multiple words but not the other way around. The following figure illustrates directional alignment models applied to the example in figure 5.4:

$$\hat{L}^D = \left\{ \begin{array}{ll} \text{no} & \text{ingen} \\ \text{one} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{very} & \text{särskilt} \\ \text{patient} & \text{mycket} \end{array} \right\}, \quad \text{link clusters}^D = \left\{ \begin{array}{ll} \text{no one} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{very} & \text{särskilt} \\ \text{patient} & \text{mycket} \end{array} \right\}$$

$$\hat{L}^I = \left\{ \begin{array}{ll} \text{no} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{very} & \text{särskilt} \\ \text{very} & \text{mycket} \\ \text{one} & \text{tålamod} \end{array} \right\}, \quad \text{link clusters}^I = \left\{ \begin{array}{ll} \text{no} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{very} & \text{särskilt mycket} \\ \text{one} & \text{tålamod} \end{array} \right\}$$

Directional link sets can be combined in several ways. The *union* of link sets ($\hat{L}^{\cup} = \hat{L}^D \cup \hat{L}^I$) usually causes many overlaps and, hence, very large link clusters. On the other hand, an *intersection* of link sets ($\hat{L}^{\cap} = \hat{L}^D \cap \hat{L}^I$) removes all overlaps and leaves only highly confident one-to-one word links behind. Using the same example from above we obtain the following alignments:

$$\hat{L}^{\cup} = \left\{ \begin{array}{ll} \text{no} & \text{ingen} \\ \text{one} & \text{ingen} \\ \text{one} & \text{tålamod} \\ \text{is} & \text{visar} \\ \text{very} & \text{särskilt} \\ \text{very} & \text{mycket} \\ \text{patient} & \text{mycket} \end{array} \right\}, \quad \text{link clusters}^{\cup} = \left\{ \begin{array}{ll} \text{no one} & \text{ingen tålamod} \\ \text{is} & \text{visar} \\ \text{very patient} & \text{särskilt mycket} \end{array} \right\}$$

$$\hat{L}^{\cap} = \left\{ \begin{array}{ll} \text{no} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{very} & \text{särskilt} \end{array} \right\}, \quad \text{link clusters}^{\cap} = \hat{L}^{\cap}$$

The union and the intersection of links do not produce satisfactory results. Another possibility is a refined combination of link sets
($\hat{L}^R = \{\hat{L}^D \cap \hat{L}^I\} \cup \{L_1^R, ..., L_r^R\}$) as suggested by Och and Ney [ON00b]. In this approach, the intersection of links is iteratively extended by additional links L_r^R which pass **one** of the following two constraints:

- A new link is accepted if both items in the link are not yet aligned.
- Mapped on a two-dimensional bitext space, the new link is either vertically or horizontally adjacent to an existing link **and** the new link does not cause any link to be adjacent to other links in both dimensions (horizontally and vertically).

Applying this approach to the example, we get:

$$\hat{L}^R = \left\{ \begin{array}{ll} \text{no} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{very} & \text{särskilt} \\ \text{very} & \text{mycket} \\ \text{one} & \text{ingen} \\ \text{patient} & \text{tålamod} \end{array} \right\}, \text{link clusters}^R = \left\{ \begin{array}{ll} \text{no one} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{very} & \text{särskilt mycket} \\ \text{patient} & \text{tålamod} \end{array} \right\}$$

Another alignment approach is the *competitive linking approach* proposed by Melamed [Mel96a]. In this approach, one assumes that there are **only** one-to-one word links. The alignment is done in a "best-first" search manner where links with the highest association scores are aligned first, and the aligned items are then immediately removed from the search space. This process is repeated until no more links can be found. In this way, the optimal alignment ($\hat{L}^C$) for non-overlapping one-to-one links is found. The number of possible links in an alignment is reduced to $min(N,M)$. Using competitive linking with our example we yield:

$$\hat{L}^C = \left\{ \begin{array}{ll} \text{no} & \text{ingen} \\ \text{very} & \text{särskilt} \\ \text{is} & \text{visar} \\ \text{one} & \text{tålamod} \\ \text{patient} & \text{mycket} \end{array} \right\}, \text{link clusters}^C = \hat{L}^C$$

Another iterative alignment approach is proposed in [Tie03]. In this approach, the link $L_x^B = [s_{x_1}, t_{x_2}]$ with the highest score in the clue matrix
$\hat{C}_{all}(s_{x_1}, t_{x_2}) = max_{s_i,t_j}(C_{all}(s_i,t_j))$ is added to the set of link clusters if it fulfills certain constraints. The top score is removed from the matrix (i.e. set to

zero) and the link search is repeated until no more links can be found. This is basically a *constrained best-first search*. Several constraints are possible. In [Tie03] an adjacency check is suggested, i.e. overlapping links are accepted only if they are adjacent to other links in one and only one existing link cluster. Non-overlapping links are always accepted (i.e. a non-overlapping links creates a new link cluster). Other possible constraints are clue value thresholds, thresholds for clue score differences between adjacent links, or syntactic constraints (e.g. that link clusters may not cross phrase boundaries). Using a best-first search strategy with the adjacency constraint we obtain the following alignment:

$$\hat{L}^B = \left\{ \begin{array}{ll} \text{no} & \text{ingen} \\ \text{very} & \text{särskilt} \\ \text{very} & \text{mycket} \\ \text{one} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{patient} & \text{mycket} \\ \text{patient} & \text{tålamod} \end{array} \right\}, \text{link clusters}^B = \left\{ \begin{array}{ll} \text{no one} & \text{ingen} \\ \text{is} & \text{visar} \\ \text{very patient} & \text{särskilt mycket tålamod} \end{array} \right\}$$

None of the alignment approaches described above produces the manual reference alignment in our example using the given clue matrix. However, simple iterative procedures come very close to the reference and produce acceptable alignments even for multi-word units, which is promising for an automatic clue alignment system. Directional alignment models depend very much on the relation between the source and the target language. One direction usually works better than the other, e.g. an alignment from English to Swedish is better than Swedish to English because in English terms and concepts are often split into several words whereas Swedish tends to contain many compositional compounds. Symmetric approaches to word alignment are certainly more appropriate for general alignment systems than directional ones.

Association measures as clues

Using association measures for estimating alignment clues lead to another complication, the proper transformation of association scores into association probabilities. The definition of estimated clues using arbitrary association measures assumes a linear correlation between the scores of the measure and the probability of the corresponding association. For example, the degree of similarity of two strings by means of LCSR assumed to be linearly correlated with the probability of the two strings being cognates. The correlation factor is expressed in the value of $w_i = \lambda_i / Z_i$ where Z_i is used to normalize the association score (i.e. to transform the score into a value between 0 and 1)

and λ_i is used to weight the score according to the correlation between the measure and the corresponding probability.

The normalization factor can be found by setting $Z_i = max_{x,y}(A_i(x,y))$, for instance $Z_{LCSR} = max_{x,y}(LCSR(x,y)) = 1$ (for identical strings x and y). However, for measures like point-wise mutual information, such a maximum is not defined for arbitrary arguments x and y. In the case of point-wise mutual information it is possible to make Z_I dependent on specific items s and t, as the maximum of point-wise mutual information is given with $max_x(I(x,t)) = max_x\left(log_2 \frac{P(x,t)}{P(x)P(t)}\right) = log_2 \frac{P(x)}{P(x)P(t)} = -log_2 P(t)$ and similarly $max_y(I(s,y)) = -log_2 P(s)$ [MS99]. Normalizing point-wise mutual information can then be done by setting $Z_{I,s,t} = max(max_y(I(s,y)), max_x(I(x,t))) = max(-log_2 P(s), -log_2 P(t))$. Other measures like t-scores may even produce negative values that can be interpreted as statistical evidence for items to avoid each other [CGHH91]. A simple way of handling negative scores is to disregard them. However, the value of general t-scores can also be arbitrarily large. The maximum of a t-score given certain items s and t also depends on the estimation of the standard error, which involves the size of the corpus as one parameter. Normalization in these cases is very inefficient because it depends on several parameters and has to be re-calculated for individual word pairs. An approximate solution for normalization is to use a large value $\tilde{Z}_i \gg A_i(s,t)$ for highly correlated items s and t. This approximation is sufficient as long as the score does not exceed $\tilde{Z}_i$.

Another important issue is the adjustment of an appropriate weighting scheme. Weights have to resemble the importance of specific clues. However, an optimization of weights has not been carried out in the present study.

In [Tie03], word alignment experiments are presented using two clues, which have been derived from association scores: cognate clues using LCSR scores and co-occurrence clues using the Dice coefficient. LCSR and Dice produce values between 0 and 1, thus, simplifying the normalization. The clue weights have been set uniformly to 0.5. One of the advantages of the clue alignment approach is that clues may even refer to overlapping multi-word units. This makes it possible to compute association scores for various words and word groups from the corpus. Two approaches to the identification of MWU candidates have been tested in [Tie03]: Pre-defined chunks using shallow parsers and arbitrary bi- and tri-grams[3].

Learning clues

As mentioned earlier, clues can be learned from aligned data. Learning clues is based on the assumption that an association clue C_j between lexical items is linearly correlated with the likelihood of f_t and f_s being features of aligned

[3] A short list of stop words has been used to recognize common phrase boundaries.

source and target language items. In this way, word alignment relations are represented as *generalized* clues from aligned training data using *clue patterns* that define the features that are assumed to carry the relational information.

One way to estimate the correlation between features of aligned items is to use a conditional probability $P(f_t|f_s)$, which is the likelihood of the features f_t of target language items given the features f_s of aligned source language items. This likelihood can be estimated using the frequency $freq(f_s, f_t)$ of co-occurring features in aligned training data, the feature frequencies $freq(f_s)$ and $freq(f_t)$ in the same data, and a maximum likelihood estimation:

$$C_j(s,t) = w_j A_j(f_s, f_t) = w_j P(f_t|f_s) \approx w_j \frac{freq(f_s, f_t)}{freq(f_t)} \tag{5.1}$$

The correlation factor is expressed as the weight w_j. Features can be any kind of information attached to the lexical items or to any contextual data. A conditional probability is "directional", i.e. estimations of $P(f_t|f_s)$ may be very different to estimations of $P(f_s|f_t)$ especially for features with very different occurrence frequencies. One possibility of a symmetric estimation of the association between features is to use the Dice coefficient for combining both conditional probabilities.

$$C_j(s,t) = w_j A_j(f_s, f_t) = w_j \frac{2 * P(f_s, f_t)}{P(f_s) + P(f_t)} \approx w_j \frac{2 * freq(f_s, f_t)}{freq(f_s) + freq(f_t)} \tag{5.2}$$

The Dice coefficient is in fact the harmonic mean of the two conditional probabilities $P(f_s|f_t)$ and $P(f_t|f_s)$ as shown in equation 2.2.

Unfortunately, word aligned training data is not available and therefore, conditional feature probabilities cannot be estimated directly. However, other clues such as association measure clues may be used to create word alignments as described earlier. The idea now is to use automatically aligned data as training data for the estimation of feature clues. In this way, alignment clues can be found in a *boot-strapping procedure*, starting with a basic alignment and learning new clues from (noisy) alignments that have been produced using previous clues. Such *self-learning* techniques are known to increase noise. However, experimental results show that the alignment gains from dynamically learned clues, as is described in [Tie03] and section 5.3. The main reason for the success of this method is that previously unused information and contextual dependencies can be integrated in the alignment process. For example, relations between part-of-speech tags can be found, which can be used to generalize the translation relation between words that belong to certain word classes. Translational dependencies between word positions can be learned and syntactic relations can be identified. Many features are possible for the estimation of dynamic clues. Four dynamic clue patterns have been defined in [Tie03]:

POS : Part-of-speech (POS) tags are used as features. This clue pattern assumes relations between POS tags of corresponding lexical items.

POS coarse : Some POS tags (e.g. SUC-tags[4] for Swedish) include detailed morphosyntactic information. *POS coarse* assumes associations between reduced tag-sets.

phrase : Chunk labels (representing common phrase types) can be used as a feature. Associations are assumed between certain chunk types that may help to find translation relations between lexical items belonging to these chunks.

position : The feature in this clue pattern is the word position of a translated item relative to the word position of the original item in the source language segment.

Many more examples than the ones above could be shown. There is a large variety of possible feature sets and many possible combinations among them. Some other examples of dynamic clue patterns are listed in table 5.2.

name	pattern (source and target)	features
lex	#text	word itself
lexpos	#text pos	word + its part-of-speech tag
postrigram	left:pos pos right:pos	part-of-speech tag trigram (previous word, current word, next word)
posposition	pos relative position	part-of-speech tag + word position relative to the position of the aligned word
chunktrigram	chunk:left:type chunk:type chunk:right:type	chunk label trigram (this chunk, previous chunk, next chunk)
chunktrigrampos	chunk:left:type chunk:type chunk:right:type pos	chunk label trigram (this chunk, previous chunk, next chunk) + part-of-speech tag

Table 5.2: Dynamic clue patterns.

[4]SUC is the Stockholm-Umeå corpus of 1 million running Swedish words [EKÅ92].

Clue patterns do not have to be symmetric like the ones described here. A clue pattern can also refer to features of different complexity for source and target languages. However, it is important to bear in mind that the design of clue patterns is important for the reliability of clues that have been learned using these patterns. Learning clues assumes that associations between translated words can be generalized as relations between certain features. As mentioned earlier, clues can be misleading, which is certainly true for dynamic clues that may "over-generalize" relations, i.e. they may express relations that do not correlate with associations we are looking for. Another crucial assumption is the independence assumption between clues, which is needed for the combination of clues. Overlapping associations indicators violate this assumption, which may influence the approach in a negative way.

Statistical alignment models and clue alignment

Word alignment clues may be derived from any source that "promises" to find associations between words and phrases. Alignment approaches used in statistical machine translation are such a source producing, among other things, bilingual lexical translation probabilities. SMT uses refined translation models for estimating lexical probabilities that are directly applicable as a resource in the clue alignment system. SMT alignment models include various kinds of dependencies, which makes them a valuable resource, evident in the experimental results (section 5.3.2). There are several results of SMT alignments, beside the lexical probabilities, that can be applied in clue alignment. SMT itself produces a word aligned bitext suitable for training. Type links can be extracted directly from the aligned corpus and may be applied as lexical resources in clue alignment. Distortion parameters can be used as an additional clue. Even fertility parameters and mappings to the "empty word" could be added to improve the clue alignment approach. However, the main resource that has been used in our experiments is the set of lexical translation probabilities, which can be directly included in clue alignment experiments.

Declarative clues

Declarative clues are pre-defined and originate from language sources such as bilingual dictionaries or contain expert knowledge such as pre-defined relations between certain feature pairs. Declarative clues have not been used in the experiments presented in [Tie03]. Further experiments (see section 5.3.3) show that simple declarative clues improve alignment results. An example of

a simple declarative clue for aligning English and Swedish texts is the part-of-speech clue in figure 5.5[5].

```
# features (source): { 'pos' => '^(..).*$/$1' }
# features (target): { 'pos' => '^(([^N].).*\Z|(N).*\@(.).*\Z)/$2$3$4' }
TO VB           V@              # TO + VERB -> VERB
DT NN           ND              # DET + NOUN -> DEFINITE NOUN
DT NN NN        ND              # DET + NOUN + NOUN -> DEFINITE NOUN
DT JJ NN        DF AF ND        # EMBEDDED ADJECTIVES IN DEFINITE NP's
DT JJ NN        DF AQ ND        #              - " -
DT JJ NN NN     DF AF ND        #
DT JJ NN NN     DF AQ ND        #
DT JJ JJ NN     DF AQ AQ ND     #
DT JJ JJ NN     DF AQ ND        #
DT JJ JJ NN     DF AF AF ND     #
DT JJ JJ NN     DF AF ND        #
```

Figure 5.5: Declarative part-of-speech clues.

Similarly, declarative clues can be defined between chunk labels, relative word positions, common function words etc. The following declarative clues have been defined for the experiments in section 5.3:

POS : Pre-defined relations between part-of-speech labels. An example for English-Swedish can be seen in appendix B.

POS-MWU : Relations between part-of-speech label sequences as shown in figure 5.5

Chunk : Pre-defined relations between chunk-labels. An English-Swedish example is included in appendix B.

Negative clues

The clues discussed above are all used as "positive" indicators for associations between lexical items. It has already been mentioned before that certain clues may be misleading depending on the reliability of the source from which they have been derived. Some clues may indicate exactly the opposite of positive clues, i.e. high clue values may signal a "negative" relation between items such as negative t-scores indicate words which "avoid" each other in a corpus. Such *negative clues* (C^-) could be utilized in alignment in order to prevent incorrect links. One way would be to use the complement of a negative clue ($1 - C^-$) in the existing framework. However, this does not correspond to the

[5]The first column contains English part-of-speech tags and the second column contains parts of Swedish part-of-speech tags which are produced by the substitution pattern which is shown in the header of the example (*target*).

actual function of a negative example as it does not reduce already existing association scores but only improves scores for items that do not match the example. A proper integration of negative clues would be a change in the clue alignment model in order to allow the reduction of association scores according to the values of negative clues. One could, for example, define the total clue as the joint probability of positive clues and the complement of negative clues. Using the independence assumption for alignment clues the total clue could be computed as the following product: $C_{all} = C^{+} * (1 - C^{-})$. However, in this way, negative clues would have a very strong influence on the indicated associations as they affect all scores. Consequently, negative clues have to be chosen very carefully in order to lead the alignment program to the correct decisions. Strong negative clues could be, for example, pre-defined links of words that should absolutely not be aligned. Examples of such links would be pairs of non-related function words, which tend to obtain rather large association scores because they co-occur frequently. Other examples would be word pairs that are very similar but are in fact "false friends". The impact of negative clues has not yet been tested and a proper way of integrating them in the system has not yet been found. It would be interesting to investigate this further in the future.

5.2 Evaluation metrics

In section 2.3.4, two types of measures for word alignment evaluation have been presented, the *ARCADE* measures for translation spotting and the *SPLIT* measures for exhaustive word-to-word alignments. Another measure has been proposed by the author of this thesis in [AMST00]. The measures in this approach (the *PWA* measures) are tailored toward gold standards which include complex MWU links of sampled words from the bitext. In the following I will review this approach, revise the evaluation measures, and propose refined metrics for word alignment evaluation (the *MWU* measures).

5.2.1 The PWA measures

For the computation of the PWA measures, the following *partiality value Q* is calculated for partially correct link proposals[6] for each reference link of the gold standard.

$$Q_x = \frac{|aligned^x_{src} \cap correct^x_{src}| + |aligned^x_{trg} \cap correct^x_{trg}|}{max(|aligned^x_{src}|, |correct^x_{src}|) + max(|aligned^x_{trg}|, |correct^x_{trg}|)}$$

[6]Partially correct links include at least one correct source language word and at least one correct target language word, i.e. $|aligned^x_{src} \cap correct^x_{src}| > 0$ and $|aligned^x_{trg} \cap correct^x_{trg}| > 0$. In all other cases the link is called *incorrect* and $Q_x \equiv 0$ by definition.

The set of $aligned^x_{src}$ includes all source language words of **all** proposed links if at least one of them is partially correct with respect to the reference link x from the gold standard. Similarly, $aligned^x_{trg}$ refers to all the proposed target language words. $correct^x_{src}$ and $correct^x_{trg}$ refer to the sets of source and target language words in link x of the gold standard. Using the partiality value Q, we can defined the recall and precision metrics as follows:

$$R_{pwa} = \frac{\sum_{x=1}^{X} Q_x}{|correct|}, P_{pwa} = \frac{\sum_{x=1}^{X} Q_x}{|aligned|}$$

The value of $|aligned|$ refers to the number of links x for which there is at least one link proposal by the system, i.e. the number of links x for which there is a link proposal n with $|aligned^n_{src} \cap correct^x_{src}| > 0$ or $|aligned^n_{trg} \cap correct^x_{trg}| > 0$. The value of $|correct|$ is the size of the gold standard (X). The measures above take the correct portions of partially correct links into account. A consequence of this definition is that partially correct links will be scored as completely correct if they cover a MWU link completely and nothing else than words of the MWU link. For example, if the reference link represents the linked MWUs "United Nations" and the Swedish translation "Förenta nationerna" the following sets of links will be considered to be completely correct with respect to the reference:
(United → Förenta, Nations → nationerna),
(United Nations → Förenta nationerna),
but also (United → nationerna, Nations → Förenta).
The first two link sets are perfectly acceptable whereas the third set is unlikely to be accepted in isolation of the MWU alignment. This problem is a consequence of the fact that the gold standard does not give any information about how the internal parts of a MWU link should be covered. However, cases like the constructed one above are usually very rare and acceptable with respect to the underlying MWU link.

Another problem with the partiality measure Q is that its value is the same for both precision and recall. Hence, precision and recall are always identical in cases where at least one partially correct link has been proposed for each reference link, i.e. $|aligned| = |correct|$.

5.2.2 The MWU measures

In order to capture the difference between precision and recall I propose the following refinements of the PWA measures for precision and recall using MWU links[7]:

[7] Again, $Q_x \equiv 0$ for incorrect links for both, precision and recall.

$$Q_x^{precision} = \frac{|aligned_{src}^x \cap correct_{src}^x| + |aligned_{trg}^x \cap correct_{trg}^x|}{|aligned_{src}^x| + |aligned_{trg}^x|}$$

$$Q_x^{recall} = \frac{|aligned_{src}^x \cap correct_{src}^x| + |aligned_{trg}^x \cap correct_{trg}^x|}{|correct_{src}^x| + |correct_{trg}^x|}$$

$$R_{mwu} = \frac{\sum_{x=1}^{X} Q_x^{recall}}{|correct|}, P_{mwu} = \frac{\sum_{x=1}^{X} Q_x^{precision}}{|aligned|}$$

These measures will be called the *MWU measures* for word alignment evaluation in the following. Let us review the example from table 2.1 on page 27 in order to illustrate the differences between the evaluation metrics which have been introduced so far.

Gold standards		alignment
complex alignment	MWU splitting	proposed links
no one → ingen	is → visar (S)	patient → tålamod
is patient → visar tålamod	is → tålamod (P)	very → mycket
very → mycket	patient → visar (P)	
	patient → tålamod (S)	
	no → ingen (S)	
	one → ingen (S)	
	very → mycket (S)	
precision	recall	F
$P_{split} = 2/2 = 1.00$	$R_{split} = 2/5 = 0.40$	$F_{split} \approx 0.57$
		$AER \approx 0.43$
$P_{arcade} = \frac{0+1+1}{3} \approx 0.67$	$R_{arcade} = \frac{0+1/2+1}{3} = 0.50$	$F_{arcade} \approx 0.57$
$P_{pwa} = \frac{0+2/4+2/2}{2} = 0.75$	$R_{pwa} = \frac{0+2/4+2/2}{3} = 0.50$	$F_{pwa} = 0.60$
$P_{mwu} = \frac{0+2/2+2/2}{2} = 1.00$	$R_{mwu} = \frac{0+2/4+2/2}{3} = 0.50$	$F_{mwu} = 0.67$

Table 5.3: Word alignment evaluation metrics.

Table 5.3 contains different evaluation scores given the links from the sample gold standard[8] (column 1 and 2) and two link proposals by an imaginary alignment system (column 3). The differences between precision scores are quite large. The example illustrates the dependence of the evaluation on particular metrics, which have been tailored towards specific alignment purposes and styles of gold standards. Note that the precision measure P_{split}

[8]Two links in the MWU-splitting gold standard have been marked as probable (P) for the sake of explanation. This might not exactly meet the guidelines for creating such references.

is only possible for *completely aligned* gold standard documents. Otherwise it is impossible to judge how many alignments the system proposes for existing gold standard links. Compare this to a word sampling method for creating gold standards. Using only samples of words makes it impossible to say if a proposed link is to be counted as incorrect with respect to the sampled link or just as another link that is not to be considered in the evaluation. The difference between the PWA measures and the MWU measures can be mainly seen in precision. In our example, the PWA measure "punishes" the proposed links because of their partiality in both, precision and recall. However, they are not necessarily wrong but incomplete. The MWU measures capture this fact by a clear difference between precision and recall. In our experiments, we applied word sampling gold standards with complex MWUs. Hence, the MWU measures are a natural choice for evaluation of the alignment performance.

5.3 Experiments and results

A large number of experiments were carried out mainly using the corpus data from the PLUG project. Results have been evaluated using the gold standards created in the same project [MA99]. An extensive summary of alignment results, using the greedy word aligner described in section 5.1.2, can be found in [AMST99][9]. Results from initial experiments with the clue alignment approach are presented in [Tie03].

In this section I will present alignment experiments which were carried out recently on three sub-corpora from the PLUG corpus, the novel "To Jerusalem and back: A personal account" by Saul Bellow in English and Swedish (Bellow), the Scania 1995 corpus in Swedish and English (Scania95en), and the Scania 1995 corpus in Swedish and German (Scania95de). These three sub-corpora were chosen for comparing alignment results for different text types (literary and technical text) and for different language pairs (Swedish-English, English-Swedish and Swedish-German). We will concentrate on the clue alignment approach as it allows many interesting combinations of techniques and resources.

First, I will describe the setup of the presented experiments. This includes a brief overview of the gold standards which have been used for evaluation. Secondly, I will present experiments using basic alignment clues. Thirdly, I will discuss the effect of declarative and dynamic clues on alignment performances. Then, I will compare different search strategies and their impact on alignment quality. Finally, I will discuss the overall alignment performance in comparison with previously achieved results.

[9] Note that the evaluation metrics used in this report differ from the ones used in this thesis.

5.3.1 Experimental setup

The three corpora that have been chosen are fairly small and therefore well suited for extensive experiments using a large number of settings. Some characteristics are presented in table 5.4.

The clue alignment approach was chosen as the basic methodology. A large number of settings have been explored and results using these settings are presented in the following sections.

The first series of experiments comprises word alignment attempts using so called "basic clues" referring to alignment clues that have been derived directly from the bitext using empirical techniques such as association measures and statistical alignment. Two types of basic clues will be distinguished: *Base1* clues refers to word associations that have been found using simple association measures based on co-occurrence and string similarity. *Base2* clues refers to basic clues, including the translation probabilities that are produced by statistical alignment.

The second series of experiments comprises alignment attempts, in which declarative clues have been added to basic clues from the section before. Three types of declarative clues have been applied as described in section 5.1.3. Different combinations of basic and declarative clues are investigated on the example of aligning one of the three test corpora, the Bellow corpus. Declarative clues have been added to both types of basic clues, base1 and base2.

The third series of experiments comprises word alignment attempts using dynamic clues as described in table 5.2 in section 5.1.3. Dynamic clues are learned from previously aligned corpora. Aligned training data are produced by the clue aligner using basic and declarative clues. The impact of the quality of training data on dynamic clues is investigated as well as the performance of the aligner when adding learned clues to basic and declarative clues. Experiments have been carried out for all three test corpora. Results are presented and discussed.

In the fourth series of alignment experiments the effect of different search strategies on the alignment performance is investigated. Seven strategies have been introduced in section 5.1.3. Their impact on aligning the Bellow corpus is discussed on the example of three different clue alignment settings.

Finally, the overall performance of the clue aligner is presented for each of the three test corpora in comparison with results which have been yielded by the Uppsala Word Aligner (UWA) and the statistical alignment tool GIZA++[10]. Six types of alignment settings are used to summarize the achieved results:

[10]GIZA++ implements IBM's translation models 1 to 5 and is freely available from http://www-i6.informatik.rwth-aachen.de/web/Software/GIZA++.html provided by Franz Josef Och. The system implements several refinements of the statistical alignment models discussed in section 2.3.2 [ON00b].

Word alignment using base1 clues, base2 clues, base1 and base2 clues in combination with declarative clues, and, finally, a combination of the latter two with dynamic clues. For each type the best result in terms of F-values has been chosen.

The gold standards

Within the PLUG project, gold standards were created for a majority of subcorpora of the project corpus. They were produced using a word sampling method and the PLUG Link Annotator (PLA) [MAA02]. 500 randomly sampled words were chosen for texts from three different genres (literary, technical, political documents) and two different language pairs (Swedish-English and Swedish-German). Gold standards include "fuzzy" links, "null" links, and complex MWU links. They were produced by several annotators using PLA and detailed guidelines [Mer99a]. The average agreement between annotators was measured for the creation of some of the gold standards and yielded about 92% [AMST99]. Figure 5.6 shows some examples of links in the gold standard of the English-Swedish Bellow corpus[11].

regular link:	call on → ta i bruk
source:	Mr Kedourie doubts that he needed "to **call on** the resources of American political science for such lessons in tyranny?
target:	Mr Kedourie betvivlar att han behövde "**ta** den amerikanska statsvetenskapens resurser **i bruk** för sådana lektioner i tyranni.
null link:	what →
source:	This is **what** has held the Jews together for thousands of years.
target:	Den gemensamma bönens band är mycket starkt.
fuzzy link:	unrelated → inte tillhör hans släkt
source:	And though he is not permitted to sit beside women **unrelated** to him or to look at them or to communicate with them in any manner (all of which probably saves him a great deal of trouble), he seems a good-hearted young man and he is visibly enjoying himself.
target:	Och fastän han inte får sitta bredvid kvinnor som **inte tillhör hans släkt** eller se på dem eller meddela sig med dem på något sätt (alltsammans saker som utan tvivel besparar honom en mängd bekymmer) verkar han vara en godhjärtad ung man, och han ser ut att trivas gott.

Figure 5.6: Links from a gold standard.

The same gold standards have been used throughout the experiments. How-

[11] The figure illustrates a typical problem with manual alignment of word samples. The example of the "fuzzy link" may seem to be odd because it does not include "to him" in the English part of the link corresponding to the Swedish translation "inte tillhör hans släkt" [does not belong to his family]. Alignment decisions have to be made and mistakes are always possible. Probably this particular alignment would have been different if the complete sentence was to be aligned instead of the sampled source language word ("unrelated") only.

ever, the clue aligner and UWA use different formats. Several links have been lost in the automatic conversion from the old format to the clue aligner format due to tokenization differences. However, the majority of links could be converted without difficulties. Table 5.4 summarizes the sub-corpora and their corresponding gold standards that have been used to evaluate the experiments presented here.

corpus				gold standard	
name	languages	type	size (words)	size	converted
Bellow	English-Swedish	literary	138,000	500	468
Scania95en	Swedish-English	technical	385,000	500	483
Scania95de	Swedish-German	technical	337,000	500	468

Table 5.4: Test corpora and gold standards.

5.3.2 Basic clues

Basic clues are directly derived from parallel corpora without word aligned training material or external knowledge sources. The following basic clue types have been used in the experiments:

Dice: The Dice coefficient for co-occurring words and multi-word units (MWUs). The threshold for Dice scores was set to 0.2 and word pairs that co-occur only once were discarded.

I: Point-wise mutual information. A threshold of 4 was used and word pairs with a co-occurrence frequency of 1 were removed.

$t-score$: The t-test as association measure. The score threshold was set to 1 and the co-occurrence frequency threshold was set to 2 as for Dice and point-wise mutual information scores.

LCSR: String similarity scores using the longest common sub-sequence ratio. The threshold was set to 0.4. Weighted string similarity measures have not been used.

GIZA: Translation probabilities derived from GIZA++. The system was applied using its standard settings (training scheme: model 1 + HMM + model 3 + model 4).

$GIZA^{-1}$: GIZA++ implements directional statistical alignment models. $GIZA^{-1}$ refers to the translation probabilities produced when aligning a bitext in the reverse direction.

I have used a very simple method for normalizing and weighting scores. *Dice* and *LCSR* scores have been weighted by a factor of 0.05. I and $t-score$ clues have been weighted by 0.005 and 0.01, respectively[12]. The *GIZA* clues are considered to be more reliable than, for instance, *Dice* clues because they are estimated with consideration of contextual dependencies. Therefore, a slightly higher weight of 0.1 is used for the *GIZA* clue. The same applies to the $GIZA^{-1}$ clue. Table 5.5 illustrates several alignment experiments using different sets of basic clues for the three bitexts.

clues	Bellow			Scania95en			Scania95de		
base1 clues	P	R	F	P	R	F	P	R	F
lcsr	38.0	29.7	33.3	46.4	29.1	35.8	60.4	32.6	42.4
I	64.0	40.2	49.4	52.9	63.4	57.7	53.8	55.5	54.7
t-score	52.3	58.4	55.2	49.9	64.6	56.3	54.4	65.1	59.3
dice	**73.7**	60.1	**66.2**	**66.8**	61.8	**64.2**	**68.7**	62.1	65.2
dice+I+t-score	58.3	62.3	60.2	53.1	65.9	58.8	59.4	66.7	62.8
dice+I+t-score+lcsr	59.3	**68.0**	63.3	55.1	**68.5**	61.1	62.3	**70.6**	**66.2**
base2 clues	P	R	F	P	R	F	P	R	F
giza	78.0	76.5	77.3	**81.9**	**82.0**	**82.0**	78.9	79.0	79.0
giza+lcsr	73.7	78.1	75.8	74.7	79.4	77.0	75.8	77.6	76.7
giza+dice	77.8	79.3	78.5	71.3	75.1	73.1	75.5	79.2	77.3
giza+dice+lcsr	75.7	**81.0**	78.3	72.4	81.0	76.5	74.9	**79.7**	77.2
$giza^{-1}$	**79.6**	80.3	**79.9**	81.7	76.4	79.0	**79.8**	79.7	**79.7**

Table 5.5: Basic word alignment clues.

A large difference can be observed between the alignment results using different sets of basic clues. Generally, a simple string matching clue is clearly the worst indicator for word alignment even for closely related languages such as Swedish, German and English. GIZA, on the other hand, is the best single clue of our basic clues. Alignments using these probabilities as alignment clues yield the highest precision among all settings for all three bitexts.

Combinations of co-occurrence clues such as the Dice clues, point-wise mutual information clues and t-score clues do not improve the overall performance when measured in terms of F-values. Such combinations generally yield higher recall at the expense of precision. However, for all three corpora, the loss in precision is so high that the overall performance drops significantly for the combination of the three measures in comparison with the single best measure among them (Dice). This can be explained by the fact that closely related clues such as co-occurrence clues violate the independence assumption made when combining them. However, a combination of the probabilistic GIZA clue and the Dice co-occurrence clue produces a slight

[12]Weights are chosen intuitively rather than using empirical investigations. Scores are simply truncated in cases where they exceed 1.

improvement for the Bellow corpus. The loss in precision is compensated by a clear increase in recall. This is not the case for the technical Scania corpora. Here, the precision decreases considerably without noticeable improvements in recall. The recall rate goes even down significantly in case of the Swedish-English Scania bitext.

Surprisingly, the LCSR clue based on string similarity does not add much to the aligner. On the contrary, performance drops in most of the cases when adding string similarity clues. This was unexpected especially in the case of technical Swedish-German text. A small improvement can be observed when adding LCSR scores to Dice based alignment clues when aligning the Scania bitexts. However, the improvement is very little and does not justify the expensive task of computing string matching clues.

The overall picture of basic clues and their influence on alignment results is very consistent even though the bitexts represent two different genres and include two different language pairs. Therefore, we may conclude that probabilistic clues yield reliable results independent of the text genre and possibly even independent of the language pair under consideration. However, these claims are too strong to be drawn from our data as they include only three related languages and two types of text. The experiments prove that probabilistic clues can be successfully applied to a variety of languages and text types. Furthermore, they imply that string matching techniques do not work well for word alignment at least according to the present investigations. However, adjustments in the weighting scheme may change this behavior.

5.3.3 Declarative clues

Declarative clues are pre-defined relations as described in section 5.1.3, i.e. they are static and independent of the particular bitext. Three declarative clue sets for English and Swedish bitexts have been tested in the present study; part-of-speech clues in form of pairs of part-of-speech tags (*pos*), a second set of part-of-speech clues based on pairs of tag sequences that match multi-word units (*pos-mwu*), and a set of chunk label pairs (*chunk*). The *pos-mwu* clues are shown in figure 5.5. The other two clue sets are listed in appendix B. Declarative clues can be combined with any other set of clues. In this study, the three sets have been combined with some of the basic clues which have been presented in the previous section. Naturally, declarative clues apply only to bitexts that include language pairs matching the defined clue features and which include necessary data such as part-of-speech tags and chunk labels. We concentrate our investigations on the Bellow corpus. The English part of the corpus has been tagged and parsed using the open-source library for natural language processing *Grok* [Bal02]. The Swedish portion has been tagged using the *TnT tagger* [Bra00, Meg02] and parsed using Beáta Megyesi's

context-free grammar parser for Swedish [Meg02]. All markup has been done automatically and no corrections have been made. Hence, errors are to be expected in the markup. The declarative clues have been weighted uniformly. Figure 5.7 illustrates the results produced when combining declarative clues with different sets of basic clues.

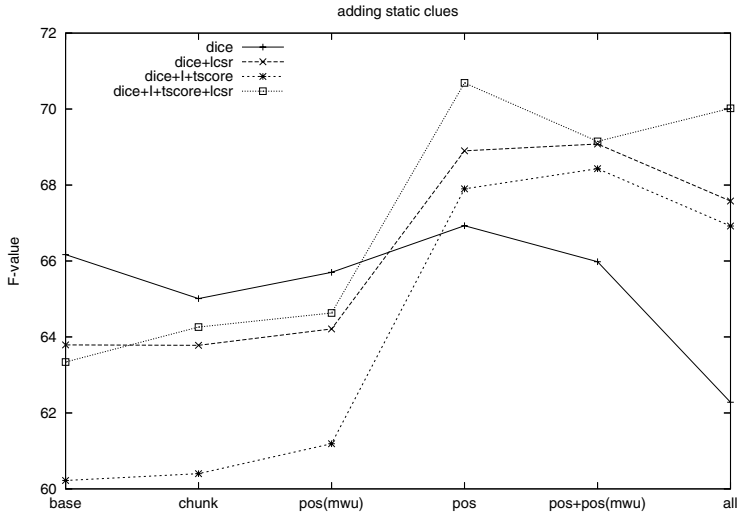


Figure 5.7: Adding declarative clues to base1 clues: pairs of part-of-speech tags (pos), pairs of part-of-speech sequences for multi-word units (pos-mwu), pairs of chunk labels (chunk).

The effect of declarative clues on the alignment performance depends very much on the set of basic clues to which they are added. Surprisingly, combinations with low performances in their basic settings (such as the combinations of three co-occurrence measures and string similarity clues (dice+I+tscore+lcsr)) outperform the basic alignment with the best performance (dice) when combined with declarative clues such as the *pos* clues. This indicates the importance of an appropriate weighting scheme, which has not been optimized in our experiments. The clue patterns seem to be in a better "balance" for the complex clue combinations in comparison with less complex ones. The declarative clues are very strong generalizations of relations between words and phrases and their influence on alignment is probably too strong in cases of less complex alignment clue sets. Similar behavior can be observed when combining declarative clues with base2 clues as shown in figure 5.8.

It is also interesting to observe that the performance decreases in most cases where all available clues are combined together. The performance drop is due

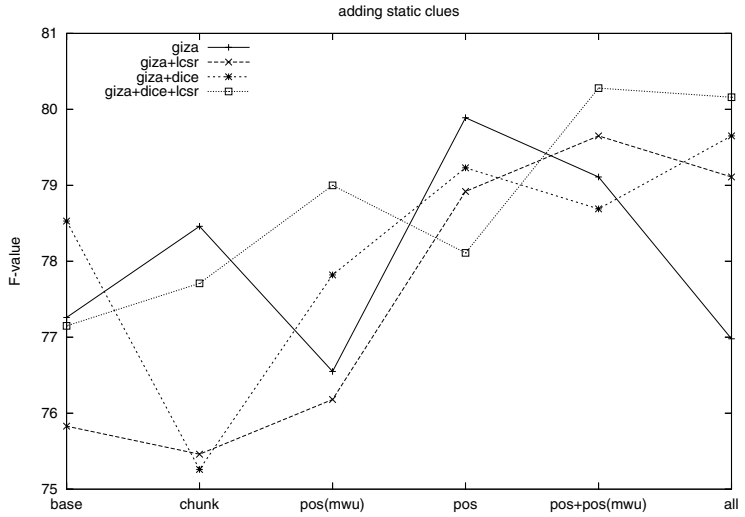


Figure 5.8: Adding declarative clues to base2 clues: pairs of part-of-speech tags (pos), pairs of part-of-speech sequences for multi-word units (pos-mwu), pairs of chunk labels (chunk).

to lower precision values that are not compensated by their corresponding recall scores. Generally, the *pos* clues seem to add the most valuable link indications to the word alignment procedure. The performance increase is mainly due to a great improvement in terms of recall. This is the case for both configurations, base1 plus declarative clues and base2 plus declarative clues.

5.3.4 Dynamic clues

A large variety of dynamic clues can be learned from previous alignments (see section 5.1.3, table 5.2). Furthermore, they can be combined in many ways. In table 5.6, alignment results for several settings of dynamic clues are presented. The quality of the training data is decisive for the performance of dynamic clues, i.e. the use of different basic clues for producing aligned material affects the set of dynamic clues that can be learned. The scores in table 5.6 illustrate the influence of training data on the performance of the word aligner using dynamic clues. Two different basic alignments have been used in order to learn several types of dynamic clues from the Bellow corpus, i.e. alignments using the Dice coefficient (Dice) and alignments using the GIZA clue. As shown in table 5.5 the Dice clues perform worse than GIZA clues. The difference amounts to more than 11% in terms of F-values (about 66% for Dice clues and more than 77% for the GIZA clues). In other words, both alignments represent

Bellow – dynamic clues (F-values)									
basic clues	p3	c3	c3p	eacl	lexp	lex	pp	p3p	lex+p3p+c3p
dice (66.2)	46.3	45.6	52.2	59.7	63.2	66.8	64.5	61.9	**74.7**
giza (77.3)	48.3	45.4	58.1	60.2	69.7	73.1	65.9	66.5	**80.3**

Table 5.6: Dynamic word alignment clues learned from two different basic clues, Dice and GIZA. Dynamic clues: POS trigram (p3), chunk trigram (c3), chunk trigram+POS (c3p), POS+POS coarse+chunk+position (eacl), wordform+POS (lexp), wordform (lex), POS+position (pp), POS trigram+position (p3p), and combinations of them (lex+p3p+c3p).

very different training sets. As expected, dynamic clues learned from these two link sets reflect the difference in that way, that clues learned from the GIZA alignment generally perform better than clues learned from Dice alignments. However, the performance differences are relatively small when considering the large difference between the two training sets. This shows that valuable alignment clues can be learned even from small and noisy alignments as the ones produced by the the Dice clues. Precision seems to be the main factor for inferring generalized dynamic clues whereas recall is a secondary aspect.

Alignment clues that involve features such as parts-of-speech and phrase types (p3, c3, c3p, eacl, pp, p3p) behave very similar for both experimental settings because they generalize relations between words and phrases on a morphological and syntactic level rather than on the lexical level. These generalizations seem to be quite similar for both training sets if one considers the resulting word alignments. The experiments in table 5.6 also show that certain generalizations are better suited for word alignment than other ones. The combination of part-of-speech tags and relative positions (*pp* and *p3p*) seem to be the best clue patterns on a morphosyntactic level for our test corpus. Other *morphosyntactic clues* perform much worse on their own (p3, c3, c3p) indicating that they do not capture much useful information for word alignment. However, a combination of very simple morphosyntactic clues as in the *eacl* setting performs much better than similar clues on their own.

On the other hand, learned clues including lexical features (*lex, lexp*) perform very differently. For them, the amount of training data (i.e. recall of the base alignment) seems to be more decisive for their alignment quality than for the others. These *lexical clues* perform significantly better when learned from the giza-alignment than from the dice-alignment.

Finally, a combination of lexical and morphosyntactic clues (*lex+p3p+c3p*) clearly out-performs all the other dynamic clue settings.

Adding dynamic clues

Table 5.7 summarizes alignment results for different bitexts and clue sets. Basic clues (as indicated in the first column) were used to create aligned training data for learning dynamic clues listed in the remaining columns. Each column contains the performances of the word aligner when adding the corresponding dynamic clues to the set of basic clues. The top scores are indicated in bold style for each row in the table.

	Bellow – adding dynamic clues (F-values)								
basic clues	p3	c3	c3p	eacl	lexp	lex	pp	p3p	lex+p3p+c3p
dice (66.2)	67.3	68.8	68.5	72.7	66.7	67.9	74.1	71.9	**74.2**
dice+lcsr+static (67.6)	68.1	70.1	70.3	72.9	70.5	71.6	74.5	72.7	**77.1**
giza (77.3)	75.7	76.8	78.3	79.8	78.4	78.4	80.7	**81.2**	81.1
giza+lcsr+static (79.1)	76.4	78.3	79.1	79.3	78.7	78.1	**82.2**	81.9	81.3
	Scania95en – adding dynamic clues (F-values)								
basic clues	p3	c3	c3p	eacl	lexp	lex	pp	p3p	lex+p3p+c3p
dice (64.2)	67.3	69.3	69.3	70.5	66.1	65.5	70.8	70.7	**72.7**
dice+lcsr+static (67.3)	68.1	69.5	68.9	68.5	68.9	70.0	71.6	71.6	**72.6**
giza (82.0)	77.8	79.8	80.3	76.7	82.5	**82.9**	81.4	82.3	81.3
giza+lcsr+static (79.3)	77.7	78.1	79.4	76.6	80.1	80.8	80.6	80.8	**83.0**
	Scania95de – adding dynamic clues (F-values)								
basic clues	p3	c3	c3p	eacl	lexp	lex	pp	p3p	lex+p3p+c3p
dice (65.2)	67.8	61.1	68.5	73.1	67.2	67.0	68.9	68.8	**70.3**
dice+lcrs+static (72.8)	72.4	69.8	72.9	72.8	**74.2**	73.5	73.2	73.8	73.9
giza (79.0)	76.4	70.9	75.4	76.8	78.5	78.0	75.5	78.4	76.5
giza+lcsr+static (78.6)	77.7	76.3	76.7	76.5	79.1	78.9	77.4	**79.3**	78.2

Table 5.7: Adding dynamic clues to different basic clues (Dice, LCSR, GIZA, all declarative clues (static)). Dynamic clues as in table 5.6.

In general, dynamic clues that do not perform well on their own do not perform well in combination with basic clues either. Another general observation is that basic clues with low alignment qualities are "easier" to improve than clues with high alignment qualities. Most of the dynamic clues that have been added to base1 clues (dice, dice+lcsr+static) boost the alignment performance for all three test corpora. On the other hand, only some dynamic clues help to improve alignments based on base2 clues (giza, giza+lcsr+static). However, certain improvements can be observed for all three test corpora when adding dynamic clues even to base2 clues. The only exception is the giza clue alignment of the Swedish-German Scania95 corpus, which did not improve in combination with any dynamic clue. In most cases, dynamic clues that perform well on their own produce the largest improvements when combined with basic and declarative clues. In several cases, the difference between the best score and the second best score is very low and a general conclusion about the quality of specific

dynamic clues in comparison with others cannot be drawn from this result. In the present experiments the possibilities of incorporating learned clues for word alignment have been investigated. The results prove that it is possible to improve word alignment results using such clues. An optimized weighting scheme is needed to get the best out of each learned clue when combined with others.

5.3.5 Different search strategies

In section 5.1.3 (pp 57) several alignment search strategies have been discussed. The clue aligner implements different strategies in order to test their impact on the alignment performance. In figure 5.9, the alignment results for three clue settings and the Bellow corpus are presented for the seven search strategies described earlier.

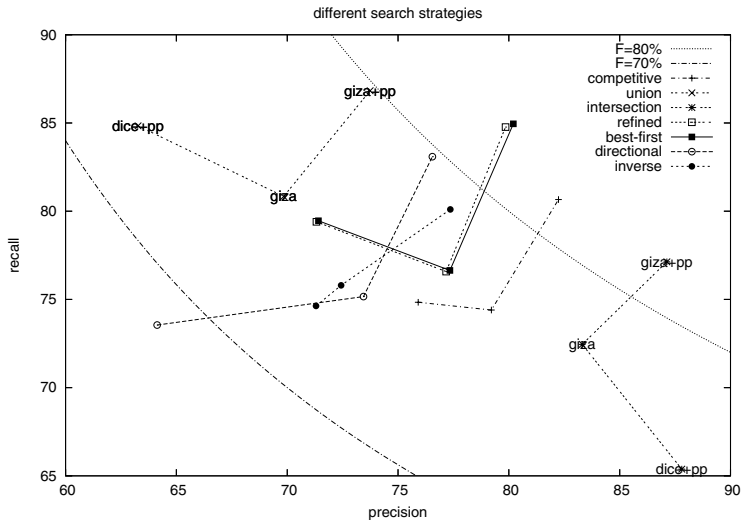


Figure 5.9: Different alignment search strategies. Corpus: Bellow, Swedish-English. Clue types: Giza++ lexicon (giza), Dice, POS + position (pp). Alignment strategies: *directional* (L^D), *inverse* directional (L^I), bi-directional *union* ($L^\cup$), bi-directional *intersection* ($L^\cap$), bi-directional *refined* (L^R), *best-first* (L^B) and competitive linking (L^C).

The figure illustrates the relation between precision and recall when using different algorithms. As expected, the intersection of directional alignment strategies yields the highest precision at the expense of recall, which is generally lower than for the other approaches. Contrary to the intersection, the union of directional links produces alignments with the highest recall

values but lower precision than all other search algorithms. Directional alignment strategies generally yield lower F-values than other refined symmetric alignment strategies. The differences between the two alignment directions are surprisingly inconsistent. Competitive linking is somewhat in between the intersection approach and the two symmetric approaches, "best-first" and "refined". This could also be expected as competitive linking only allows non-overlapping one-to-one word links. The refined bi-directional alignment approach and the constrained best-first approach are almost identical in our examples with a more or less balanced relation between precision and recall.

According to figure 5.9, alignment strategies can be chosen to suit particular needs. Concluding from our experiments, restrictive methods like the intersection approach or competitive linking should be chosen if results with high precision are required (which is mostly found among one-to-one word links). This is, for example, the case in automatic extraction of bilingual lexicons where noise should be avoided as much as possible. Other strategies should be chosen for applications, which require a comprehensive coverage as, for example, machine translation.

5.3.6 UWA, clue alignment and SMT

In the previous sections, the clue alignment approach has been explored using a variety of different settings and resources. One of the main advantages of this approach is its modularity, which makes it possible to integrate many different alignment resources. In this section we will look at the clue alignment results in comparison with other alignment approaches, i.e. the greedy word aligner implemented in the Uppsala Word Aligner (UWA) and the statistical alignment approach implemented in the GIZA++ toolbox. We will look at the three test corpora which have been used throughout our investigations separately.

We will use the following name conventions to refer to different settings: The best result for a clue alignment using basic clues only, except GIZA clues, are denoted *base1*. The best result of an alignment with basic clues including the GIZA clues are called *base2*. Alignments including declarative clues contain the string "+*stat*" in their labels and alignments including dynamic clues include the string "+*dyn*". Identical names do not necessarily refer to the same setting for each of the three corpora. There are usually several settings that fit into the given category. Here, we will always use the best of our results for the given category and each specific corpus, which is not necessarily taken from results presented in previous sections. Note that all clues are weighted as described in the previous sections. An optimization of weights for the combination of clues has not been carried out. Alignment results using GIZA++ are denoted with *GIZA++* and refer to standard settings of the GIZA++ system (IBM model 1 + HMM + IBM model 3 + IBM model

4, five iterations for each model). *GIZA++ inverse* refers to alignments in the opposite translation direction. The best result using the Uppsala Word Aligner is denoted *UWA*.

Figure 5.10 illustrates clue alignment results for different settings together with alignment performances of UWA and GIZA++ for the English-Swedish Bellow corpus.

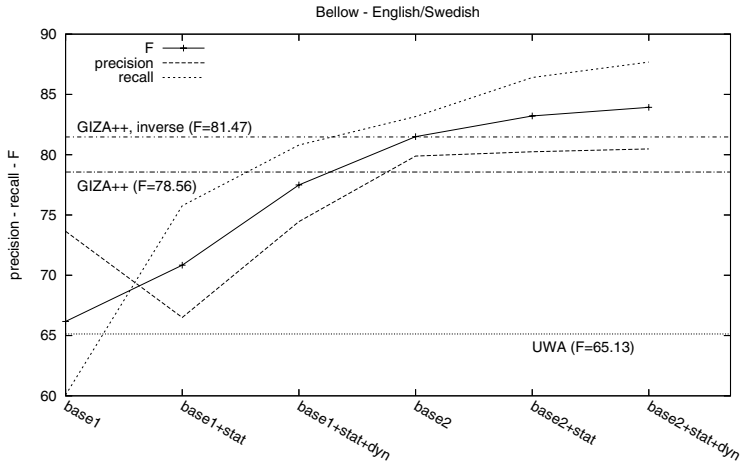


Figure 5.10: Comparison of alignment settings for the Bellow corpus.

The clue aligner clearly outperforms the simple UWA alignment approach even for basic setups. The main improvement can be found in recall, which is much higher than for the restrictive UWA approach. UWA yields high scores in terms of precision (about 86%) but low recall scores (only 52% in our example) resulting in an F-value of about 65% as indicated in figure 5.10. The precision of the clue aligner is below 81% for all settings but with corresponding recall values way above 52%. The clue alignment results are generally below the GIZA++ alignments as long as GIZA clues are not included in the alignment procedure. However, using the GIZA++ dictionaries implies a major improvement of the aligner performance and, combined with declarative and dynamic clues, it goes beyond the GIZA++ alignment result. The top result with an F-value of over 83.9% is clearly an improvement compared to the other approaches including GIZA++.

Figure 5.11 illustrates alignment results for the Swedish-English Scania 1995 corpus.

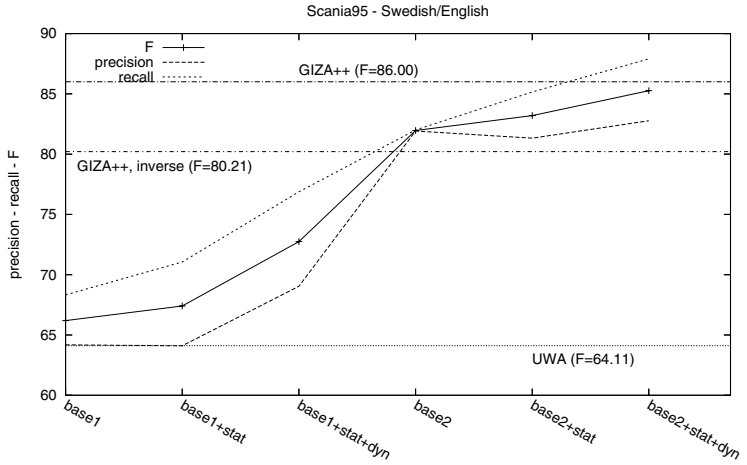


Figure 5.11: Comparison of alignment settings for the Swedish-English Scania 1995 corpus.

GIZA++ performs very well on the Swedish-English Scania corpus. Both, precision and recall are high above 80% (86.15% in precision and 85.85% in recall). Swedish and English are two very closely related languages and technical manuals such as the documents in the Scania corpus are very consistent in their terminology. Statistical alignment approaches profit from consistency as it can be seen in the excellent alignment that is produced by GIZA++. Certainly, the alignment direction suits the statistical aligner. GIZA++ links Swedish words to each English word in the corpus, which makes it possible to align several English words to the same Swedish word. The GIZA++ alignment in the other direction does not reach the same performance (only about 80.21%). This is consistent with the results of the alignments of the Bellow corpus. The highest performance there is also obtained when aligning Swedish words to their English counterparts. However, the performance difference between the two GIZA++ alignments is even larger for the Scania corpus than for the Bellow corpus. Similarly to the alignments of the Bellow corpus, GIZA++ outperforms the clue alignment approach when using base1 clues only. Using the GIZA dictionary (base2 clues) improves the clue alignment significantly. However, it does not reach the performance of the statistical alignment even though it touches its result (85.26% for a combination of base2, static and dynamic clues). This may be explained by the importance of distortion and fertility parameters estimated by GIZA++ but not used in the clue aligner. Scania documents contain many

short sentences and sentence fragments with a large number of compound terms. Many alignment problems can be solved with the help of positional dependencies and relations between compositional compounds (in Swedish) and non-compositional compounds (in English).

Finally, in figure 5.12 alignment results for the Swedish-German Scania 1995 corpus are illustrated.

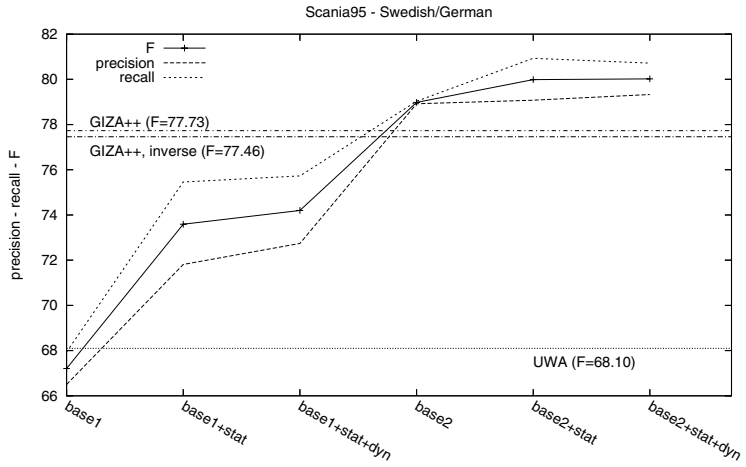


Figure 5.12: Comparison of alignment settings for the Swedish-German Scania 1995 corpus.

Similarly to the other two bitexts, the clue alignment performance increases when adding clues to the process and outperforms the Uppsala Word Aligner. Basic alignments using association measures (base1 clues) are worse than the alignments produced by GIZA++. However, adding the GIZA clue to the aligner increases the performance beyond the result of the statistical alignment. An interesting fact is that the base2 clue alignment performs better than both GIZA++ alignments. The difference between the two directional alignments performed by GIZA++ is minimal. However, symmetric alignment approaches, such as the clue alignment strategy used here, outperform directional strategies. The relation between Swedish and German is different from the relation between Swedish and English. Compounding is common in both languages (Swedish and German) and, therefore, the impact of the alignment direction is not very great. The quality of the statistical alignment is much lower than for the Swedish-English bitext, probably due to syntactic and morphological differences between Swedish and German, which are larger than between Swedish and English. It would be interesting to investigate these differences in detail by means of alignment examples.

5.4 Summary and discussion

This chapter includes one of the main contributions of the thesis. Word alignment techniques have been discussed in detail and their applications to different data collections have been investigated. Two word alignment systems have been presented, the Uppsala Word Aligner (UWA), with its iterative alignment strategy, and the matrix-based Clue Aligner. Both aligners have a modular design and integrate several alignment resources and techniques. UWA (section 5.1.2) implements a knowledge-poor approach using association measures based on co-occurrence and string similarity. It combines empirical data with alignment heuristics in an iterative linking procedure. Multi-word units can be handled in two ways: They can be identified either by means of n-gram statistics prior to the word linking step or dynamically within the linking procedure. The Clue Aligner (section 5.1.3) is a modular word alignment tool that supports the probabilistic combination of various resources. It is based on so-called alignment clues, which indicate translational relations between words and phrases according to their associated features. These features may be any set of linguistic annotations, word position, contextual information, and the words and phrases themselves. Alignment clues can be derived from association statistics (as in UWA), from statistical alignment models, and from aligned training data (dynamic clues). They can also be pre-defined (declarative clues). Dynamic clues can be learned even from noisy training data such as automatically word-aligned bitexts.

Both systems mentioned above have been evaluated using manually produced reference alignments (so-called gold standards). Fine-grained metrics for precision and recall have been defined for this purpose (section 5.2). UWA and the Clue Aligner have been applied to three different bitexts from two different genres (technical and literary texts). They include two language pairs, Swedish-English (in both translation directions) and Swedish-German (only in that direction). Different settings of the Clue Aligner have been tested and results have been compared to UWA alignments and to statistical alignments using the external toolkit GIZA++ (section 5.3). The experiments have demonstrated that the Clue Aligner is superior to UWA for all three test corpora. It has also been shown that the clue alignment approach can be used to improve statistical alignments produced by, for instance, GIZA++. The performance of the Clue Aligner depends on the quality of the alignment clues and on the weighting scheme used for the combination of clues. The optimization of the clue aligner settings and weights remains to be investigated.

6 Applications of extracted bilingual data

In this chapter I will briefly summarize four investigations on applications of lexical data that have been extracted from parallel corpora using our alignment techniques described in the previous chapter. Two studies concern the field of computational lexicography and terminology, and two the field of machine translation. Section 6.1 discusses the use of word aligned data in monolingual lexicography. Section 6.2.1 presents a prototype for interactive translation using a translation predictor, and section 6.2.2 describes work on the enhancement of a machine translation prototype for industrial use using corpus based translation data.

6.1 Computational lexicography and terminology

Parallel corpora were discovered as a resource for lexicography years ago. The main interest was focused on cross-lingual lexicography. Bilingual term extraction techniques have been widely explored using word alignment techniques as described in the previous chapters. Later on, lexicographers discovered parallel data as a valuable resource even for monolingual investigations. Parallel corpora have been successfully used for automatic word sense disambiguation and for the discovery of semantic relations as mentioned in section 2.4. The next two sections briefly describe two studies in monolingual lexicography. Further details can be found in two publications by the author of this thesis [Tie01d, Tie01a].

6.1.1 Morphological and semantic relations

Word alignment can be used to create bilingual translation lexicons (collections of word type links) of the vocabulary included in parallel corpora. There are usually a large amount of link alternatives for most of the automatically aligned words and phrases. These link alternatives are processed with the aim of the automatic discovery of morphological and semantic relations between them. It is assumed that word alignment alternatives are related to each other in some sense, i.e. that they are inflectional variants of each other, that they share a semantic meaning, or that they are translations of homonymic or polysemous

words. The approach described in [Tie01d] uses simple filters with different settings in order to find relational categories between link alternatives.

Initially, word alignment results are filtered using simple stop word lists, frequency thresholds, string type filters, and simple stemming functions. Then, string similarity measures are used to divide word type links with translation alternatives into different categories. Three string similarity measures are used for this purpose:

- the longest common sub-sequence ratio normalized by the length of the longer string (*LCSR*)
- the longest common sequence of contiguous characters normalized by the length of the shorter string (*MATCH*)
- the edit distance, i.e. the minimal number of deletions, insertions, and substitutions needed to transform one string into the other (*EDIT*)

The differences between the three measures are illustrated in figure 6.1.

```
MATCH     r e g u l a t e          MATCH  a n r u f e n
      i r r e g u l a r i t i e s         a n r i e f e n
  LCS     r e g u l a     t   e      LCS  a n r u   f e n
```

$LCSR(regulate, irregularities) = 8/14 \approx 0.57$
$MATCH(regulate, irregularities) = 6/8 = 0.75$
$EDIT(regulate, irregularities) = 6$

$LCSR(anrufen, anriefen) = 6/8 = 0.75$
$MATCH(anrufen, anriefen) = 3/8 = 0.375$
$EDIT(anrufen, anriefen) = 2$

Figure 6.1: String matching techniques.

Three categories are investigated [Tie01d]: inflectional relations, morphological relations, and semantic relations.

Filter one focuses on the first category, i.e. inflectional relations between alternative translations in word type links. Inflectional variants are assumed to be very similar in spelling, with a limited number of edit operations required to change one variant into another. Therefore, the LCSR measure was used with a high threshold (0.75) in combination with a low edit distance (< 5). The LCSR measure is chosen because it captures similar ties of non-contiguous character sequences. Infix modifications such as the vowel change from [u] in the German "anrufen" to [ie] in its past tense form "anriefen" are very common. The MATCH measure fails to capture this as illustrated in figure 6.1. The following two groups of Swedish translations, which have been aligned to the English words "often" and "fault code", are typical examples of word type links that have been categorized as inflectional variants using filter one:

English	Swedish	English	Swedish
fault code	felkod	often	oftast
	felkoder		ofta
	felkoden		oftare
	felkoderna		
	felkodernas		
	felkodens		

Table 6.1: Inflectional relations among link alternatives.

The second filter focuses on category two: morphological relations such as derivation and compounding. Morphological variants are assumed to contain a common root but to differ from each other more than inflectional variants. Therefore, the MATCH measure was used with a high threshold (0.75) for the identification of common roots and the threshold for the edit distance was set to a minimum of 5 operations to distinguish word pairs from category one. For example, the following two groups of Swedish translations, which have been aligned to the English words “load” and “adjustment”, have been put into category two by this filter:

English	Swedish	English	Swedish
load	belastningen	adjustment	justeras
	last		justeringen
	belastas		justeringarna
	belastad		grundjustering
	belasta		

Table 6.2: Derivational relations among link alternatives.

The third filter emphasizes semantic relations (category three). The general assumption here is that translation alternatives with semantic relations are not derived from a common root and, therefore, are significantly unlike each other. Hence, translation alternatives with a low LCSR score (below 0.5) and an edit distance above 4 are chosen. Different types of semantic relations can be found among translation alternatives using this filter. Some examples are shown below:

synonymy (Swedish)		polysemy/homonymy		homographs (English)	
English	Swedish	English	Swedish	English	Swedish
cable	ledningen	work	fungera	report	skriv
	vajer		arbeta		protokoll
	kablage	plate	skylt	test	provas
	kabel		plattor		test
reduce	minska	note	anteckna		provning
	reducerar		observera		testa
specify	specificerade				
	angivet				
	angivna				
together	ihop				
	tillsammans				

Table 6.3: Semantic relations among link alternatives.

All three filters are certainly very simple and cannot be used for fully automatic classification of link alternatives into relational categories. Parameters of each filter are adjustable. A change of parameters leads to different results. A manual evaluation of a sample of 50 outcomes/filter from the experiments on English-Swedish data is presented in [Tie01d], illustrating the overlap of relation types between word type links that have been classified by the three filters[1]:

filter	inflected	derived	compound	synonym	homonym
filter 1	86%	12%			
filter 2		16%	74%		
filter 3			14%	28%	16%

Table 6.4: Relational categories.

The study shows that simple filters can be used for a rough classification of translation alternatives into relational categories. In other words, bilingual data and word alignment techniques can be used to identify relations between words in one language. The same filters can also be used to identify possible alignment errors (in particular filter 3). The present study is based on UWA alignments. A better result especially for the last category is expected for improved word alignments, for instance, based on the clue aligner.

[1]The remaining entries are alignment errors.

6.1.2 Phrasal term extraction

Terminology often contains a large number of phrasal constructions. One of the main difficulties in automatic term extraction is the recognition of such units. The use of parallel corpora for the identification of phrasal terms in one language is investigated [Tie01a]. Domain specific parallel corpora bear a large number of terms for both languages. Word alignment techniques can be used to align terms among words of the general vocabulary. Word alignment, as discussed in the previous chapter, supports the alignment of multi-word-units (MWUs). Multi-word terms are assumed to be consistently translated into corresponding terms in other languages. In [Tie01a] it is argued that word alignment can help to improve the extraction of phrasal terms assuming that the alignment is performed with high accuracy.

The experiments presented in [Tie01a] are based on alignments using the Uppsala Word Aligner (UWA) and parallel technical texts from the Scania1998 corpus in Swedish and English. UWA uses simple statistical methods for prior identification of word collocations, which are candidates for the alignment of MWUs. Simple collocation statistics do not provide a very accurate phrase list and usually over-generate phrase unit candidates. The experiments in [Tie01a] show that word alignment eliminates a large number of incorrect candidates and also adds significant phrase candidates to the list.

Phrasal term candidates have been investigated at three different stages of the alignment process: Candidates, which have been identified using collocation statistics, candidates, which have been annotated in the corpus, and candidates, which have been aligned by the word aligner. Statistically extracted collocations include many overlapping MWUs. The annotation does not allow overlapping units, i.e. they disappear in the annotation step. MWUs are used by the word aligner in the same manner as single words. Additional MWUs derived by string similarity measures are also employed by the alignment program. The results are hard to evaluate in terms of accuracy and completeness. It is usually unknown how many phrasal terms actually are included in a corpus. Furthermore, it is not straightforward to define the nature of a correct phrasal term. [Tie01a] presents several quantitative and qualitative evaluations, which indicate the performance of the system. Evaluations have been carried out for the English part of the extracted terms.

First, an existing approved collection of terms of the same domain is used to estimate the *recall* of the phrasal term extraction. Recall gives the percentage of correct terms recognized, in comparison with the number of existing terms. Not all terms of the reference collection are actually included in the corpus and, therefore, the measure cannot be seen as an absolute value. However, it can be used to compare different stages of the extraction. It is shown in [Tie01a] that alignment does not reduce the number of recognized phrasal terms significantly, indicating that alignment does not eliminate many

of the correct phrasal terms. However, word alignment reduces the number of extracted phrasal terms by 50% compared to the list of initially extracted word collocations. This indicates an improvement of *precision*[2] given a more or less constant score in recall.

The second evaluation in [Tie01a] uses syntactic patterns, which have previously been used for term extraction in order to evaluate identified terms. Typical noun-phrase patterns are used to match extracted phrasal terms and the evaluations show that word alignment increases the portion of terms that match these patterns whereas terms, which have been eliminated, match the patterns significantly less frequently. This indicates that word alignment improves the precision of phrasal term extraction, assuming that correct terms mainly belong to the syntactic categories described by the patterns.

Finally, a manual evaluation was performed on a random sample of 900 extracted phrasal terms. The evaluation shows that the portion of accepted terms increases when annotating MWUs and aligning them in parallel texts. At the same time, a large number of terms that have been excluded by the aligner are classified as invalid. Furthermore, most of the phrasal terms added during alignment are marked as correct.

In conclusion, word alignment can be used to improve simple statistical methods for phrasal term extraction. Constraints on syntactic structures would certainly improve simple collocation statistics. However, the presented example demonstrates that word alignment can be used as a tool for the improvement of phrasal term recognizer when syntactic analyzers are not available. The result depends very much on the performance of the word aligner, which also depends on the language pair. Swedish and English fit well into the framework as Swedish and English are very different in terms of compounding and term construction. Many English multi-word terms correspond to single Swedish compounds. This can be used by the word aligner to identify links between MWUs and single words.

[2]Precision gives the percentage of correct terms among all extracted term candidates.

6.2 Translation

Machine translation and translation support are two tasks, which can gain from multilingual corpora and alignment techniques. The following section briefly presents a prototype for interactive machine translation using a corpus-based translation predictor, and thereafter, the enhancement of an existing machine translation system using corpus-derived translation data is described.

6.2.1 Translation prediction

In [Tie01b] a prototype for interactive translation is presented that applies lexical data extracted from parallel corpora. The system predicts on demand translations of unknown words when writing in a second language. It essentially works as a background dictionary with contextual disambiguation which can be invoked by the user when typing texts in a foreign language. The prototype uses automatically extracted bilingual dictionaries with translation alternatives and indexed sentence-aligned parallel corpora. Unknown words are marked with a leading question mark ('?'). Each time a new word boundary is detected, the predictor tries to replace unknown words with translations that correspond to the marked word in the current context. The program takes pre- and post-context into account in the source and the target language. The disambiguation process is very simple: Additional local context is used as long as there are translation alternatives which can be found in a similar context. The program alternates between pre- and post-context using only direct neighbors. The most frequent alternative is chosen of all translations which are found for the largest context that produces translation alternatives. The user may decide if (s)he accepts the prediction or not. In the latter case the user may continues to type additional words, which will be used as additional context when the system starts the disambiguation process next time. The prototype has a simple interface shown in figure 6.2.

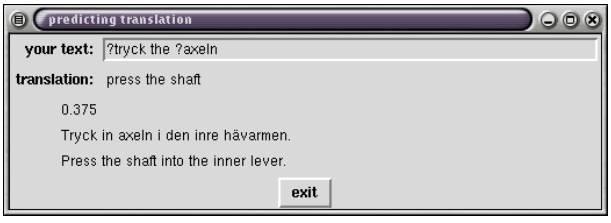


Figure 6.2: The translation predictor.

[Tie01b] describes experiments with a small-scale parallel corpus of Swedish and English technical documents. The Uppsala Word Aligner (UWA) was used to extract a bilingual dictionary from the MATS corpus with

100,000 words in Swedish and English. Half of the corpus has been used as the "training" corpus, i.e. a sentence aligned corpus of about 50,000 words has been indexed. Source language items with multiple translations in the extracted dictionary have been chosen and 611 test sentences have been extracted from the "evaluation" corpus, which comprises the other half of the MATS corpus with also about 50,000 words. Test sentences have been retrieved automatically by looking for sentences containing one of the ambiguous source language words and exactly one of its translation alternatives. An example is given below:

dictionary:	fyll → fill|charge|refill
source sentence:	**fyll** systemet med luft .
target sentence:	**charge** the system with air .
test sentence:	**?fyll** the system with air .

Now, target sentences can be used as the gold standard and test sentences can be processed by the predictor in a batch run. The results of the batch process are compared to the baseline. In the baseline, unknown words are simply replaced with the most frequent translation according to the dictionary. The experiments showed an improvement from about 71% for the baseline to about 80% correct translations using the predictor.

The translation predictor is a very simple tool for interactive translation. It can be improved in several ways. Using word forms in the index produces very sparse data. A simple stemming function may improve the results and the robustness of the system. Linguistic information could also improve the performance. Surrounding part-of-speech tags, for example, could be used for the disambiguation process. Other parameters such as word position could possibly improve the prediction accuracy. The main resource, however, is the bilingual dictionary. The system certainly needs a comprehensive dictionary. In [Tie01b], an automatically extracted dictionary has been used which covers only parts of the vocabulary and includes erroneous links. However, the results were encouraging, which promises well for larger experiments using improved alignment techniques.

6.2.2 Scaling up machine translation

The MATS system is a machine translation platform that has been developed in the MATS project, a joint project on the methodology and application of a translation system. It combines tools which have been developed in previous projects such as the Uppsala Chart Parser (UCP) and the transfer-based machine translation engine MULTRA. An overview of the system can be found in [SFT+02]. One of the main goals of the MATS project was the integration of corpus-based lexical material in order to scale up the translation prototype for

industrial use. The MATS corpus has been compiled for development purposes and the MatsLex database has been implemented for handling lexical data. Scaling up a transfer-based machine translation system involves a large effort in lexicon development. Here, the word alignment software comes into the picture. Word alignment, as described previously, produces rough bilingual wordform lexicons. These data are the starting point for the extension of the lexical database, which is one of the main resources for the translation engine. In MATS, the lexical database provides default translations of each word in the input text, morphosyntactic descriptions in terms of specific codes for recognized words and semantic features if available. Word alignment results have been lemmatized, filtered and manually checked before being entered into the database [Löf01]. The lexicon has been extended by over 7,000 lexemes in this way, which is a substantial enlargement of the tiny lexicon that existed in the original MULTRA prototype with its 59 lexemes[3].

The MATS system can be seen as a machine translation platform. It combines several modules and provides a pipe-line based architecture and common interfaces for putting modules together. The system can easily be extended by additional modules. In the future, we plan to add further data-driven methods into the system. In the on-going project KOMA corpus based machine translation is emphasized. The idea is to continue the development of the MATS system to work as a hybrid system combining transfer-based and data-driven methods. The integration of translation memories is a first step towards a hybrid translation system. Further modules may work with example-based machine translation techniques and statistical approaches to machine translation. In this way, parallel corpora play a central role in the development of the system:

- Raw lexical data are extracted from parallel corpora via word alignment.
- Aligned sentences serve as translation memory and as input for example-based translation.
- Parallel corpora are used as training material for statistical machine translation.
- Previous translations provide reference material for evaluation.

Recycling cannot be more thorough.

[3] A similar project has recently been carried out in co-operation with Systran. The lexical components of the Swedish-English and Swedish-Danish engines of the EC Systran machine translation system have been build from automatically extracted word type links using our Clue Aligner. This project is part of the European Commission Contract, SDT/MT 2003-1: Extension of EC Systran to Danish and Swedish into English.

6.3 Summary

In this chapter, four examples of applications of automatically extracted bilingual lexical data have been presented.

Word type links produced by our word aligner have been used for two applications in monolingual lexicography. The first application uses translation alternatives found by the word aligner for the identification of morphological and semantic relations between them. Simple filters are used to refer groups of alternative translations to relational categories. The second application concerns the use of word alignment for the identification of phrasal terms. Our experiments have shown that word alignment can be used to improve the quality of phrasal terminology lists, which are extracted by means of collocation statistics.

The last two applications relate to machine translation. Word alignment results are used together with a bitext index to implement a translation predictor for interactive machine translation. The prototype that has been developed can be used to find translations of words in context. The last application concerns the use of word alignment for the extension of existing machine translation systems. Raw alignment data have been used to scale up the lexical databases of two translation engines for industrial use in specific textual domains. One engine is the in-house machine translation platform MATS and the other is the EC Systran translation engine.

7 Conclusions

This final chapter contains a summary of contributions and prospects for future work.

7.1 Contributions

Four aspects of the work with translation corpora have been discussed in this thesis: The collection and compilation of parallel corpora, the implementation of tools for processing such corpora, the development and evaluation of word alignment techniques, and finally, the application of these techniques, data and tools to tasks in the field of natural language processing.

The thesis contains the following contributions:

Parallel corpora

Five parallel corpora have been built in co-operation with our project partners: PLUG, Scania1998, MATS, KOMA, and OPUS. They contain about 36 million words altogether in mainly technical documents from different domains. The main focus of investigation was set on the PLUG corpus, which also contains literary and political text. All corpora are sentence aligned. Some parts have been automatically analyzed and annotated with linguistic markup. OPUS differs from the other corpora as it contains a large number of languages in parallel (up to 60 in some parts) including many non-European languages, and that it is freely available. OPUS is also the largest of the five corpora (about 30 million words in its current version, 0.2).

Corpus processing tools and lexical databases

Several tools have been developed while working on the thesis. The Uplug system includes general corpus processing tools such as tokenizers, sentence splitters, tools for format conversion and n-gram statistics. Two word alignment systems were integrated within the framework of the Uplug toolbox, the Linköping Word Aligner (LWA) from the Department of Computer and Information Science at Linköping University and the Uppsala Word Aligner (UWA), which was developed and implemented by the author of this thesis. Another more advanced word aligner, the Clue Aligner, has been developed in the thesis. It represents a framework for the integration of empirical

and linguistic alignment resources within the Uplug environment. Tools for the automatic evaluation of alignment results are integrated in both systems. Uplug including UWA and LWA is freely available for research purposes. The Clue Aligner can be run via a web interface (UplugWeb), which includes several tools for processing monolingual and bilingual text corpora. External tools such as part-of-speech taggers can also be used via UplugWeb. Furthermore, web-based concordance tools have been implemented for our parallel corpora.

The MatsLex database uses a set of tools developed in the thesis for the storage of lexical data in a relational database management system. MatsLex can be used in a multi-user environment with tools for querying and updating the database. It includes web interfaces and command line interfaces. Lexical data for the machine translation system MATS are extracted from MatsLex via scripts that have been tailored towards the system such that the lexical database and the translation system together form a "glass-box" development platform for machine translation.

Word alignment techniques and evaluation

Various word alignment techniques have been developed in the thesis. UWA applies a "greedy" word and phrase alignment method using iterative linking cascades. It combines association measures and alignment heuristics. One of the association measures is based on string matching algorithms. Algorithms for the automatic construction of weighted character matching functions have been developed and presented in the thesis. One of the main contributions is the clue alignment approach. It allows the combination of many resources for the alignment of words and phrases in parallel corpora. Alignment clues can be derived from association measures, statistical alignments and linguistic resources. They can also be learned from aligned training data. It has been demonstrated that alignment clues can be combined to improve word alignment results and that dynamic clues can be learned even from noisy training data such as automatically aligned corpora. The clue aligner provides a framework for the combination of alignment resources and a test-suite for the comparison of word alignment strategies. Both, UWA and the Clue Aligner, include tools for the automatic evaluation of alignment results. Evaluation techniques have been discussed in the thesis in detail. Refined evaluation metrics for word alignment have been defined for the purpose of comparison.

Word alignment experiments

The presented alignment techniques have been applied to a variety of data and tasks. The word alignment approaches have been tested on several bitexts and language pairs. Chapter 5 contains detailed discussions of word alignment experiments, which have been carried out for three sub-corpora of the PLUG

corpus representing two different genres (technical and literary texts) and including two language pairs (Swedish-English and Swedish-German). The experiments show that the clue alignment approach is superior to alignments by means of UWA and that it is state of the art in comparison with alternative alignment systems. The main advantage of the clue aligner is that new resources can easily be integrated into the system. In this way, external resources such as results of other alignment tools can be used by the system to improve its performance.

Applications of extracted bilingual lexical data

Several applications of alignment tools and parallel corpora to natural language processing have been explored in the thesis. In two studies the use of parallel data for monolingual lexicography has been investigated. It has been demonstrated that word alignment can be used to identify morphological and semantic relations between words and to improve the recognition of multi-word terms. Another field of study that has been investigated is machine translation. An approach to interactive translation using a translation predictor has been examined. It applies automatically extracted bilingual dictionaries and their links back to parallel corpora, from where they originate, for the prediction of word translations in context. Extracted translation dictionaries are also used in the transfer-based machine translation system MATS and the lexical component of the Swedish-English and Swedish-Danish EC Systran translation systems.

7.2 Future work

Many aspects of parallel corpora as a source for tasks in natural language processing remain to be explored further. One obvious extension for work in the near future is the application of presented alignment techniques to other language pairs. The OPUS corpus represents an excellent source for such work.

Many future tasks are related to the OPUS project. The parallel corpus in this project is meant to grow further. The extension of the corpus is already in progress; additional multilingual document collections are in the queue to be processed and added to the corpus. One of our goals for the near future is to use web crawlers to collect parallel data automatically from the internet. Another task is to add linguistic markup to the corpus for many more languages. Porting tools to new languages via alignment is one possibility to obtain linguistic markup for additional parts of the corpus that will be investigated.

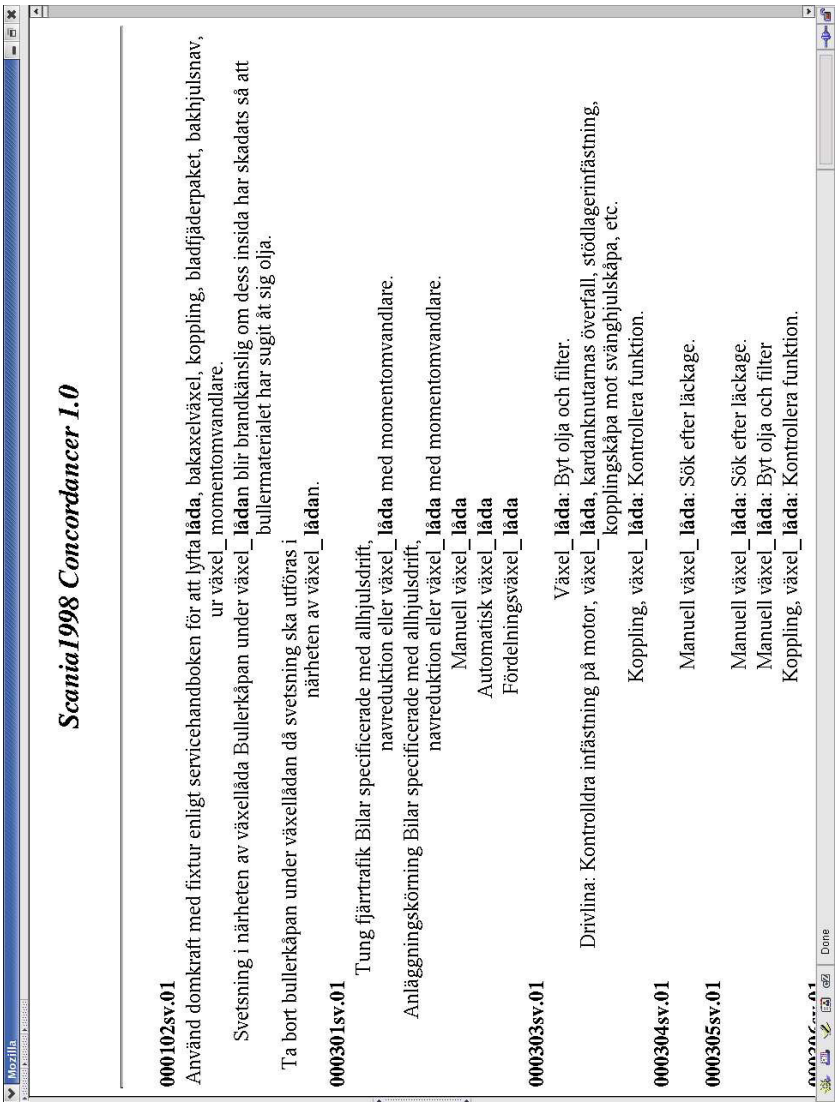
The clue aligner can be improved in many respects. System parameters have to be optimized in order to increase the performance. Weights for combining alignment clues could be learned automatically from data. Additional

resources, further contextual dependencies, and negative clues could also be explored.

A major goal for the future is the integration of alignment tools, corpus management tools, translation systems of various kinds and multilingual terminology databases into a web-based translation workbench. Such a platform could be used to build hybrid machine translation systems, on-line resources for interactive and manual translation, tools for computer-aided language learning. multi-lingual lexicography and linguistic investigations across languages.

A Screen shots

A.1 A monolingual concordance tool

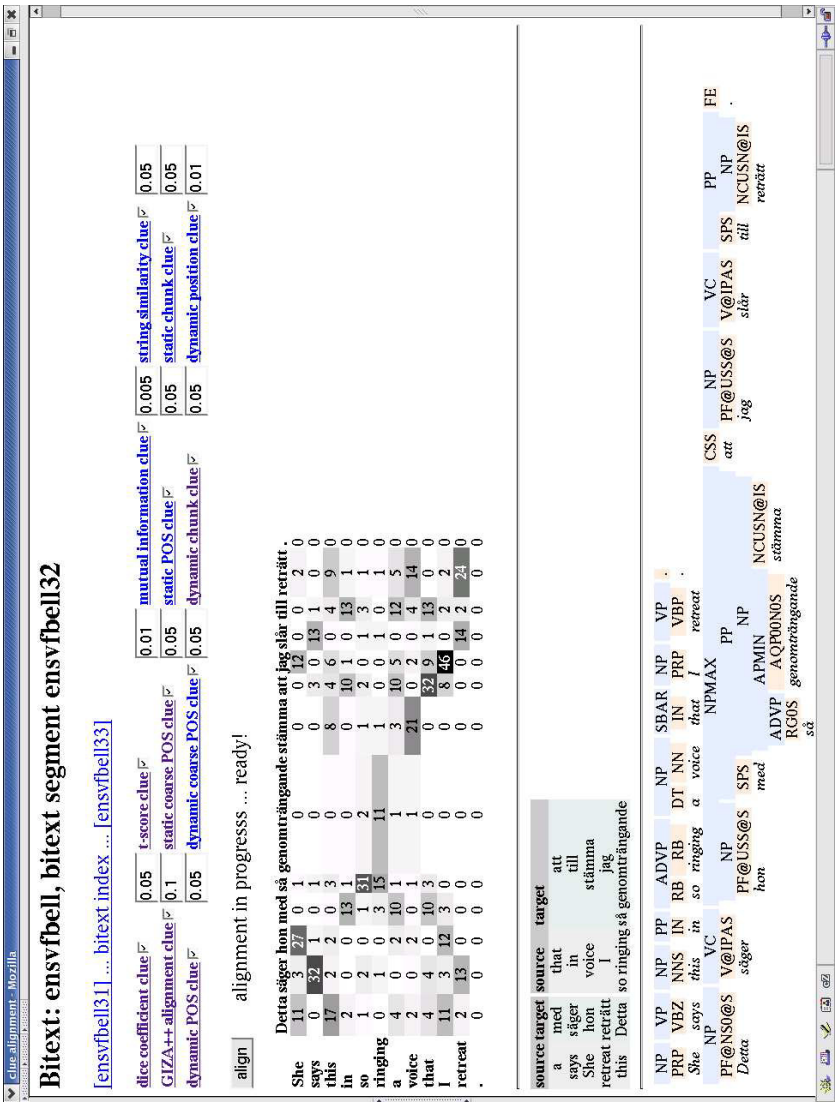


A.2 A multilingual concordance tool

	OOEN	OOFR	OODE	OOES	OOSV
122	In the text box here , enter the numerical value or select it with the up-arrow or down-arrow **buttons** .	Saisissez la valeur numérique dans la zone de texte ou sélectionnez une valeur à l' aide des boutons fléchés .	Geben Sie hier in das Textfeld den numerischen Wert ein , oder wählen Sie ihn mit der Pfeil-aufwärts- oder Pfeil-abwärts-Schaltfläche .	Introduzca en el campo de texto el valor numérico o use las teclas de flechas para seleccionarlo .	Ange det numeriska värdet i det här textfältet eller välj det med pil-upp&t - eller pil-ned&t-knapparna .
4846	On any window edge where another window is docked you will see two **buttons** which allow you to show , hide or fix the window .	Sur chaque bordure de fenêtre comportant une fenêtre ancrée se trouvent deux boutons vous permettant de masquer , d' afficher ou encore de fixer les fenêtres .	An jedem Fensterrand , an dem sich angedockte Fenster befinden , gibt es zwei Schaltflächen , mit denen Sie die Fenster ein-/ausblenden oder fixieren können .	En todos los bordes de ventana , que contengan a su vez otras ventanas acopladas , hay dos botones que permiten ocultar todas las ventanas acopladas o transformarlas en una " posición fija " .	Vid varje fönsterkant med förankrade fönster finns det två kommandoknappar med vilka du kan visa / dölja eller fixera fönstren .
22550	If the window is docked to the left of the workplace , the following **buttons** appear depending on the state :	Lorsque la fenêtre est ancrée au bord gauche de la zone de travail , vous voyez les icônes suivantes selon le mode défini :	Ist das Fenster am linken Rand des Arbeitsbereichs angedockt , sehen Sie je nach eingestelltem Zustand folgende Symbole :	Cuando la ventana está acoplada al margen izquierdo del área de trabajo , se muestran los siguientes símbolos , según el estado definido :	Om fönstret förankrats vid arbetsområdets vänstra sida visas följande symboler , beroende på vilket läge som ställts in :
23848	optional text , frames , **buttons** , etc.	Textes optionnels , cadres , boutons , etc.	optional Texte , Rahmen , Schaltflächen usw.	textos opcionales , marcos , botones etc.	texter , ramar , kommandoknappar osv .
25391	The following **buttons** are shown on the screen if you have checked the Synchronize contents check box and activated the New Doc. button .	Si vous avez coché la case Synchroniser le contenu et activé le bouton Nouveau document , un autre bouton s' affiche à l' écran :	Folgende Schaltfläche erhalten Sie auf dem Bildschirm angezeigt , wenn Sie das Markierfeld Inhalte synchronisieren markiert haben und die Schaltfläche Neues Dokument betätigen .	Los siguientes botones se muestran en la pantalla cuando se activa la casilla de verificación Sincronizar contenidos y se pulsa en el botón Nuevo documento .	Följande kommandoknappar visas på bildskärmen när Du har markerat kryssrutan Synkronisera innehåll och klickar på kommandoknappen Nytt dokument .
31398	These **buttons** start the dialog used to set printer properties .	Permet d' afficher une boîte de dialogue servant à paramétrer les propriétés de l' imprimante .	Diese Schaltfläche ruft einen Dialog auf , welcher zur Einstellung der Drucker-Eigenschaften dient .	Este botón abre un diálogo previsto para la configuración de las propiedades de la impresora .	Den här kommandoknappen öppnar en dialogruta där du kan ställa in skrivarens egenskaper .
35112	This check box changes to Including styles , as soon as you select an attribute via the Attributes or Format **buttons** .	La case à cocher intitulée Y compris les styles apparaît lorsque vous avez défini l' attribut à rechercher à l' aide des boutons Attributs ou Format .	Bei Ersetzen durch können Sie eine neue Absatzvorlage wählen , die auf Wunsch auf die gefundenen Absätze angewendet wird .	Esta casilla pasa a denominarse Buscar estilos , cuando se establece un atributo a buscar dentro del botón Atributos o Formato .	Den här rutan heter Inklusive mallar när du har definierat ett attribut för sökningen med hjälp av knapparna Attribut eller Format .
42950	To go to a specific record	Vous disposez de plusieurs	Zum Anwählen des Datensatzes in der	Para seleccionar un	Det finns flera

A.3 The UplugWeb corpus manager

A.4 Clue alignment visualization



B Declarative clues for English and Swedish

Below is a set of declarative part-of-speech clues for English-Swedish bitexts using the Penn Treebank tagset for English (first column) and the two initial characters of the morphosyntactic description (msd) from the SUC tagset for Swedish (second column).

```
# features (source): { 'pos' => undef }
# features (target): { 'pos' => '^(..).*$/$1' }
JJ      AF      # PERFECT PARTICIPLE
JJ      AP      # PRESENT PARTICIPLE
JJ      AQ      # ADJECTIVE
JJR     AQ      # ADJECTIVE (COMPARATIVE)
JJS     AQ      # ADJECTIVE (SUPERLATIVE)
CC      CC      # CONJUNCTION
IN      CS      # SUBJUNCTION
DT      D0      # DETERMINER
DT      DF      # DETERMINER
DT      DH      # DETERMINER
DT      DI      # DETERMINER
.       FE      # PUNCTUATION
IN      SP      # PREPOSITION
CD      MC      # CARDINAL NUMBER
CD      MO      # CARDINAL NUMBER
PRP     PF      # PERSONAL PRONOUN
PRP$    PS      # POSSESSIVE PRONOUN
POS     PS      # POSSESSIVE ENDING
RB      RG      # ADVERB
RB      RH      # ADVERB
RBR     RG      # ADVERB, COMPERATIVE
RBS     RG      # ADVERB, SUPERLATIVE
VB      V@      # VERB, BASE FORM
VBD     V@      # VERB, PAST TENSE
VBG     V@      # VERB, GERUND/PRESENT PARTICPLE
VBN     V@      # VERB, PAST PARTICIPLE
VBP     V@      # VERB, SINGULAR, PRESENT, NON-3RD-PERSON
VBZ     V@      # VERB, 3RD PERSON, SINGULAR
VBN     AF      # VERB, PAST PARTICIPLE <-> PERFECT PARTICIPLE
VBG     AP      # VERB, PRESENT PARTICPLE
TO      CI      # to --> att
NN      NC      # NOUN, SINGULAR --> INDEFINITE
NNS     NC      # NOUN, PLURAL --> INDEFINITE
NNP     NP      # PROPER NOUN, SINGULAR
NNPS    NP      # PROPER NOUN, PLURAL
```

The following clue set defines relations between basic chunks in English and Swedish bitexts. The labels refer to basic chunk types in English (first column) and basic phrases recognized by Beáta Megyesi's parser for Swedish [Meg02] (second column).

```
# features (source): { 'c.*:type' => undef }
# features (target): { 'c.*:type' => undef }
VP           VC
VP           INFP
ADVP         ADVP
VP ADVP VP   VC ADVP VC
VP ADJP VP   VC ADVP VC
ADJP         APMIN
ADJP         APMAX
ADJP         ADVP PP APMIN
PP           PP
PP NP        PP NP
PP NP        PP APMIN NP
NP PP NP     NPMAX
NP PP NP     NP
NP PP NP     NP PP NP
NP           NP
NP           NP APMIN NP
NP           ADVP APMIN NP
NP           ADVP NP
```

Index

Author index

References

[Abn91] Steven P. Abney. Parsing by chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*, pages 257–278. Kluwer Academic Publishers, Boston, 1991.

[Ahr99] Lars Ahrenberg. Superlinks: A new approach to containing transfer in machine translation. Technical report, Linköping University, Linköping, Sweden, 1999.

[Ahr03] Lars Ahrenberg. The KOMA corpus. Technical report, Linköping University, Linköping, Sweden, 2003.

[AMA98] Lars Ahrenberg, Magnus Merkel, and Mikael Andersson. A simple hybrid aligner for generating lexical correspondences in parallel texts. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and the 17th International Conference on Computational Linguistics (ACL/COLING)*, pages 29–35, Montreal, Canada, 1998. Morgan Kaufmann Publishers.

[AMA02] Lars Ahrenberg, Magnus Merkel, and Mikael Andersson. A system for incremental and interactive word linking. In *Proceedings from The 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 485–490, Las Palmas, Spain, May 2002.

[AMST99] Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein, and Jörg Tiedemann. Evaluation of LWA and UWA. Technical Report 15, Department of Linguistics, Uppsala University, Uppsala, Sweden, 1999.

[AMST00] Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein, and Jörg Tiedemann. Evaluation of word alignment systems. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation, (LREC)*, volume III, pages 1255–1261, Athens, Greece, 2000.

[AOCJ+99] Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John D. Lafferty, I. Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith,

and David Yarowsky. Statistical machine translation. Technical report, The Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1999.

[ARM01] Gabriel Lopes António Ribeiro and João Mexia. Extracting translation equivalents from portuguese-chinese parallel texts. *Journal of Studies in Lexicography*, 11(1):118–194, 2001.

[Arp95] Antti Arppe. Term extraction from unresctricted text. In Kimmo Koskenniemi, editor, *Proceedings of the 10th Nordic Conference on Computational Linguistics (NODALIDA)*, Helsinki, Finland, May 1995.

[AS00] Ingrid Almqvist and Anna Sågvall Hein. A language checker of controlled language and its integration in a documentation and translation workflow. In *Proceedings of the Twenty-Second International Conference on Translating and the Computer*, Translating and the Computer 22, Aslib/IMI, London, November 2000.

[Bal02] Jason Baldridge. Grok - an open source natural language processing library. http://grok.sourceforge.net/, 2002.

[BBP+94] Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, John R. Gillet, John D. Lafferty, Harry Printz, and Luboš Ureš. The Candide system for machine translation. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 157–163, Plainsboro, NJ, March 1994. Morgan Kaufman Publishers.

[BCD+90] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[BDD96] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[BDDM93] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistcal machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June 1993.

[Bel76] Saul Bellow. *From Jerusalem and back: a personal account*. The Viking Press, New York, USA, 1976.

[Bel77] Saul Bellow. *Jerusalem tur och retur*. Bonniers, Stockholm, 1977. översättning av Caj Lundgren.

[BLM91] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 169–176, Berkeley, CA, 1991.

[BM96] Chris Brew and David McKelvie. Word-pair extraction for lexicography. In *Proceedings of International Conference on New Methods in Natural Language Processing*, pages 45–55, Bilkent, Turkey, 1996.

[BMT+00] Chris Brew, David McKelvie, Richard Tobin, Henry Thompson, and Andrej Mikheev. *The XML Library LT XML version 1.2 - User documentation and reference*. Language Technology Group, Edinburgh University, Edinburgh, Scotland, 2000.

[Bor98] Lars Borin. Linguistics isn't always the answer: Word comparison in computational linguistics. In *Proceedings of the 11th Nordic Conference on Computational Linguistics (NODALIDA)*, pages 140–151, Center for Sprogteknologi and Department of General and Applied Linguistics, University of Copenhagen; Denmark, January 1998.

[Bor99] Lars Borin. Enhancing tagging performance by combining knowledge sources. In Hans Lindquist och Magnus Levin Gunilla Byrman, editor, *Korpusar i forskning och undervisning. Corpora in research and teaching. Papers from the ASLA symposium Corpora in research and teaching*, pages 19–31, Växjö University, Växjö, Sweden, November 1999.

[Bor00] Lars Borin. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th International Conference on Computational Linguistics. COLING 2000*, pages 97–103, Universität des Saarlandes, Saarbrücken, Germany, 2000.

[Bor02] Lars Borin. Alignment and tagging. In *Parallel Corpora, Parallel Worlds*, pages 207–218. Rodopi, Amsterdam, New York, 2002. Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden, 1999.

[Bra00] Thorsten Brants. TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference ANLP-2000*, pages 224–231, Seattle, WA, 2000.

[Bro96] Ralf D. Brown. Example-based machine translation in the Pangloss system. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, pages 169–174, Copenhagen, Denmark, 1996.

[Bro97] Ralf D. Brown. Automated dictionary extraction for 'knowledge-free' example-based translation. In *Proceedings of the 7th International on Theoretical and Methodological Issues in machine Translation, TMI-97*, pages 111–118, Santa Fe, NM, 1997.

[Bro00] Ralf D. Brown. Automated generalization of translation examples. In *Proceedings of the 18th International Conference on Computational Linguistics, COLING-2000*, pages 125–131, Saarbruecken, Germany, 200.

[CBO03] Chris Callison-Burch and Miles Osborne. Bootstrapping parallel corpora. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts - Data Drievn Machine Translation and Beyond*, pages 44–49, Edmonton, Canada, May/June 2003.

[CG91] Kenneth W. Church and William A. Gale. Bconcordances for parallel text. In *7th Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62, 1991.

[CGHH91] Kenneth W. Church, William A. Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In Uri Zernik, editor, *Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

[Chr94] Oliver Christ. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX)*, pages 22–32, Budapest, 1994.

[Chu93] Kenneth W. Church. Char align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 1–8, 1993.

[CKY03] Eugene Charniak, Kevin Knight, and Kenji Yamada. Syntax-based language models for statistical machine translation. In *8th International Workshop on Parsing Technologies*, Nancy, France, 2003.

[Dai95] Béatrice Daille. Combined approach for terminology extraction: Lexical statistics and linguistic filtering. Technical Report 5, Department of Linguistics, Lancaster University, 1995.

[DC94] Ido Dagan and Kenneth W. Church. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 34–40, Stuttgart, Germany, 1994.

[DCG93] Ido Dagan, Kenneth W. Church, and William A. Gale. Robust bilingual word alignment for machine-aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Respektives*, pages 1–8, Columbus, OH, 1993.

[DR02] Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 255–262, Philadelphia, PA, July 2002.

[Dun93] Ted E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[Dyv98] Helge Dyvik. A translational basis for semantics. In Stig Johansson and Signe Oksefjell, editors, *Corpora and Crosslinguistic Research: Theory, Method and Case Studies*, pages 51–86. Rodopi, 1998.

[Dyv02] Helge Dyvik. Translations as semantic mirrors: From parallel corpus to Wordnet. In *23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English (ICAME)*, Gothenburg, Sweden, May 2002.

[EKÅ92] Eva Ejerhed, Gunnel Källgren, and Ola Wennstedt Magnus Åström. The linguistic annotation system of the Stockholm-Umeå corpus project. description and guidelines. Technical report, Department of Linguistics, Umeå University, 1992.

[FC94] Pascale Fung and Kenneth W. Church. K-vec: A new approach for aligning parallel texts. In *Proceedings 15th International Conference on Computational Linguistics (COLING)*, pages 1096–1102, Kyoto, Japan, August 1994.

[FHN00] Radu Florian, John C. Henderson, and Grace Ngai. Coaxing confidences from an old friend: Probabilistic classifications from transformation rule lists. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 26–34, Hong Kong, October 2000.

[FIP96] George F. Foster, Pierre Isabelle, and Pierre Plamondon. Word completion: A first step toward target-text mediated IMT. In *Proceedings*

of 16th International Conference on Computational Linguistics (COLING), pages 394–399, Copenhagen, Denmark, 1996.

[FM94] Pascale Fung and Kathleen R. McKeown. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 81–88, Columbia, MD, 1994.

[FM97] Pascale Fung and Kathleen R. McKeown. A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation, Special Issue on New Tools for Human Translators*, 12(1-2):53–87, 1997.

[Gau98] Éric Gaussier. Flow network models for word alignment and terminology extraction from bilingual corpora. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and the 17th International Conference on Computational Linguisticss (COLING)*, pages 444–450, Montreal, Canada, 1998. Morgan Kaufmann Publishers.

[GC91a] William A. Gale and Kenneth W. Church. Identifying word correspondences in parallel texts. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, 1991.

[GC91b] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–184, 1991.

[GCY92] William A. Gale, Kenneth W. Church, and David Yarowsky. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112, Montreal, Canada, June 1992.

[GKK03] Daniel Marcu Grzegorz Kondrak and Kevin Knight. Cognates can improve statistical translation models. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pages 46–48, Edmonton, Canada, May 2003.

[Gor71] Nadine Gordimer. *A Guest of Honour*. Cape, London, 1971.

[Gor77] Nadine Gordimer. *Hedersgästen*. Bonniers, Stockholm, 1977. översättning av Magnus K:son Lindberg.

[GRCH96] Rob Gaizauskas, Pete Rodgers, Hamish Cunningham, and Kevin Humphreys. GATE User Guide. `http://gate.ac.uk/`, 1996.

[Gri98] Ralph Grishman. Tipster text architecture design. Technical report, New York University, New York, USA, October 1998.

[GVONC01] Ismael García-Varea, Franz Josef Och, Hermann Ney, and Francisco Casacuberta. Refined lexicon models for statistical machine translation using a maximum entropy approach. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 204–211, Toulouse, France, July 2001.

[Hie98] Djoerd Hiemstra. Multilingual domain modeling in Twenty-One: Automatic creation of a bi-directional translation lexicon from a parallel corpus. In Peter-Arno Coppen, Hans van Halteren, and Lisanne Teunissen, editors, *Proceedings of the 8th Meeting of Computational Linguistics in the Netherlands (CLIN)*, number 25 in Language and Computers: Studies in Practical Linguistics, pages 41–58, Nijmegen, The Netherlands, 1998. Rodopi, Amsterdam, Atlanta.

[Hol03] Maria Holmqvist. Översättningssystemet T4F - en implementation för ATIS-domänen [The translation system T4F - an implemantation for the ATIS domain]. Technical report, NLPLAB, Department of Computer and Information Science, Linköping University, Linköping, Sweden, 2003.

[IPD00] Nancy Ide and Greg Priest-Dorman. Corpus encoding standard - document CES 1. Technical report, Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, New York, USA and LORIA/CNRS, Vandoeuvre-lés-Nancy, France, 2000.

[Isa92] Pierre Isabelle. Bi-textual aids for translators. In *Proceedings of the 8th Annual Conference of the UW Centre for the New OED and Text Research*, pages 76–89, University of Waterloo, Waterloo, Canada, 1992.

[ISM03] Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. Automatic construction of machine translation knowledge using translation literalness. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 155–162, Budapest, Hungary, April 2003.

[JH94] Stig Johansson and Knut Hofland. Towards an English–Norwegian parallel corpus. In Gunnel Tottie Udo Fries and Peter Schneider, editors, *Creating and Using English Language Corpora*, pages 25–37. Rodopi, Amsterdam, Atlanta, 1994.

[JK95] J.S. Justeson and S.M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

[Kay97] Martin Kay. The proper place of men and machines in language translation. *Machine Translation*, 12:3–23, 1997. Originally appeared as a Xerox PARC Working Paper in 1980.

[KKN+94] Hans Karlgren, Jussi Karlgren, Magnus Nordström, Paul Pettersson, and Bengt Wahrolén. Dilemma - an instant lexicographer. In *Proceedings of the 15th International Conference on Computational Linguisticss (COLING)*, Kyoto, Japan, 1994. available from http://citeseer.nj.nec.com/132140.html.

[KM01] Taku Kudoh and Yuji Matsumoto. Chunking with support vector machines. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Pittsburgh, PA, USA, 2001. available from http://citeseer.nj.nec.com/kudo01chunking.html.

[KOM01] KOMA. The KOMA project. http://www.ida.liu.se/~nlplab/koma/, 2001.

[KR88] Martin Kay and Martin Röscheisen. Text-translation alignment. Technical report, Xerox Palo Alto Research Center, March 1988.

[KT90] Judith L. Klavans and Evelyne Tzoukermann. The BICORD system: Combining lexical information from bilingual corpora and machine readable dictionaries. In Hans Karlgren, editor, *Proceedings of the 13th International Conference on Computational Linguistics (COLING)*, volume 3, pages 174–179, Helsinki, Finland, 1990.

[Kup93] Julian M. Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 17–22, Columbus, OH, 1993.

[LFL00] Philippe Langlais, George F. Foster, and Guy Lapalme. TransType: A computer-aided translation typing system. In *Proceedings of Embedded Machine Translation Systems - Workshop II held in con-*

junction with NAACL/ANLP2000, pages 46–51, Seattle, Washington, May 2000.

[Löf01] Camilla Löfling. Att skapa ett lemmalexikon för manuell och maskinell översättning [To create a lemma lexicon for manual and automatical translation]. Master's thesis, Uppsala University, Uppsala, Sweden, 2001.

[LW01] Karl Lundstedt and Katarina Wahlund. Översikt av syntaktiska strukturer i MATS-korpusen [Overview of syntactical structures in the MATS corpus], 2001. Working report, Department of Linguistics, Uppsala University.

[MA99] Magnus Merkel and Lars Ahrenberg. Evaluating word alignment systems. Technical report, Linköping University, Linköping, Sweden, 1999.

[MA00] Magnus Merkel and Mikael Andersson. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proceedings from the Conference on Computer-Assisted Information Retrieval (RIAO)*, pages 737–746, Paris, France, 2000.

[MAA02] Magnus Merkel, Mikael Andersson, and Lars Ahrenberg. The PLUG link annotator - interactive construction of data from parallel corpora. In Lars Borin, editor, *Parallel Corpora, Parallel Worlds*. Rodopi, Amsterdam, New York, 2002. Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden,1999.

[MAT00] MATS. The MATS project. http://stp.ling.uu.se/mats, 2000.

[Meg02] Beáta Megyesi. Shallow parsing with POS taggers and linguistic features. *Journal of Machine Learning Research: Special Issue on Shallow Parsing*, 2:639–668, 2002.

[Mel95] I. Dan Melamed. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In David Yarovsky and Kenneth Church, editors, *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 184–198, Boston, MA, 1995. Association for Computational Linguistics.

[Mel96a] I. Dan Melamed. Automatic construction of clean broad-coverage lexicons. In *Proceedings of the 2nd Conference the Association for Machine Translation in the Americas (AMTA)*, pages 125–134, Montreal, Canada, 1996.

[Mel96b] I. Dan Melamed. A geometric approach to mapping bitext correspondence. In Eric Brill and Kenneth Church, editors, *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–12, Philadelphia, PA, 1996.

[Mel97a] I. Dan Melamed. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, Providence, 1997.

[Mel97b] I. Dan Melamed. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics (COLING)*, pages 305–312, Madrid, Spain, 1997.

[Mel97c] I. Dan Melamed. A Word-to-Word Model of Translation Equivalence. In *Proceedings of the 35th Conference the Association for Computational Linguistics*, Madrid, 1997.

[Mel98] I. Dan Melamed. Annotation style guide for the Blinker project, version 1.0. IRCS Technical Report 98-06, University of Pennsylvania, Philadelphia, PA, 1998.

[Mel00] I. Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, June 2000.

[Mer99a] Magnus Merkel. Annotation style guide for the PLUG link annotator. Technical report, Linköping University, Linköping, Sweden, 1999.

[Mer99b] Magnus Merkel. *Understanding and enhancing translation by parallel text processing*. Linköping studies in science and technology, dissertation no. 607, Linköping University, Department of Computer and Information Science, Linköping, Sweden, 1999.

[MH96] Elliott Macklovit and Marie-Louise Hannan. Line'em up: Advances in alignment technology and their impact on translation support tools. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, Montreal, Canada, 1996. available from http://citeseer.nj.nec.com/macklovitch96line.html.

[MKY+00] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. Morphological analysis system chasen version 2.2.1 manual. http://chasen.aist-nara.ac.jp/chasen/bib.html.en, December 2000.

[MNA94] Magnus Merkel, Bernt Nilsson, and Lars Ahrenberg. A phrase-retrieval system based on recurrence. In *Proceedings from the 2nd Annual Workshop on Very Large Corpora*, pages 99–108, Kyoto, Japan, 1994.

[MP03] Rada Mihalcea and Ted Pedersen. An evaluation exercise for word alignment. In *Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton, Canada, May 2003.

[MR01] Arul Menezes and Stephen D. Richardson. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of Workshop on Example-based Machine Translation at the MT Summit VIII*, pages 35–42, Santiago De Compostela, Spain, 2001.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

[Och99] Franz Josef Och. An efficient method for determining bilingual word classes. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 71–76, 1999.

[ON00a] Franz-Josef Och and Hermann Ney. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 1086–1090, Saarbrücken, Germany, July 2000.

[ON00b] Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, 2000.

[ON03] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[OTN99] Franz Josef Och, Christoph Tillmann, and Hermann Ney. Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 20–28, University of Maryland, MD, USA, 1999.

[OZN03] Franz Josef Och, Richard Zens, and Hermann Ney. Efficient search for interactive statistical machine translation. In *Proceedings of the 10th*

Conference of the European Chapter of the Association for Computational Linguistics (EACL), pages 387–393, Budapest, Hungary, April 2003.

[Rat98] Adwait Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, University of Pennsylvania, 1998.

[RLM00] António Ribeiro, José Gabriel Pereira Lopes, and João Mexia. Extracting equivalents from aligned parallel texts: Comparison of measures of similarity. In *Proceedings of the International Joint Conference IBERAMIA/SBIA*, pages 339–349, São Paulo, Brazil, November 2000.

[RM95] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In David Yarovsky and Kenneth Church, editors, *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 82–94, Boston, MA, 1995.

[RM97] Philip Resnik and I. Dan Melamed. Semi-automatic aquisition of domain-specific translation lexicons. In *Proceedings of the Conference on Applied Natural Language Processing*, Washington, D.C., 1997.

[Såg02] Anna Sågvall Hein. The PLUG project: Parallel corpora in Linköping, Uppsala, and Göteborg: Aims and achievements. In Lars Borin, editor, *Parallel Corpora, Parallel Worlds*, number 16 in Working Papers in Computational Linguistics and Language Engineering. Rodopi, Amsterdam, New York, 2002. Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden,1999.

[SAJ+99] Anna Sågvall Hein, Lars Ahrenberg, Gary Jaekel, Sören Lindh, and Barbro Atlestam. Om maskinöversättning. NUTEKs slutrapport av regeringsuppdrag. Technical Report N1999/7189/ITFOU 1999-09-22, NUTEK/Näringsdepartementet, 1999.

[SAS97] Anna Sågvall Hein, Ingrid Almqvist, and Per Starbäck. Scania Swedish - A basis for multilingual machine translation. In *Translating and the Computer 19. Papers from the Aslib conference*, London, November 1997.

[Sch94] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, September 1994. http://www.ims.uni-stuttgart.de/ schmid/.

[SFI92] Michel Simard, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 67–81, Montreal, Canada, 1992.

[SFT+02] Anna Sågvall Hein, Eva Forsbom, Jörg Tiedemann, Per Weijnitz, Ingrid Almqvist, Leif-Jöran Olsson, and Sten Thaning. Scaling up an MT prototype for industrial use - databases and data flow. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation, (LREC)*, volume V, pages 1759–1766, Las Palmas de Gran Canaria, Spain, 2002.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[Sim99] Michel Simard. Text-translation alignment: Three languages are better than two. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 2–11, University of Maryland, MD, 1999.

[SM90] Frank A. Smadja and Kathleen R. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th. Annual Meeting of the Association for Computational Linguistics*, pages 252–259, Pittsburgh, PA., 1990.

[SMH96] Frank A. Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics, 22(1)*, pages 1–38, 1996.

[Ste92] Graham A. Stephen. String search. Technical report, School of Electronic Engineering Science, University College of North Wales, Gwynedd, 1992.

[TB02] Dan Tufis and Ana-Maria Barbu. Lexical token alignment: Experiments, results and applications. In *Proceedings from The 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 458–465, Las Palmas, Spain, 2002.

[Tie97] Jörg Tiedemann. Automatical extraction of translation equivalents from aligned bilingual corpora. Master's thesis, Department of Linguistics, Uppsala University / Otto-von-Guericke-Universität Magdeburg, 1997.

[Tie98] Jörg Tiedemann. Compiling the Scania1998 corpus. Technical report, Department of Linguistics, Uppsala University, Uppsala, Sweden, 1998.

[Tie99a] Jörg Tiedemann. Automatic construction of weighted string similarity measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 213–219, University of Maryland, MD, 1999.

[Tie99b] Jörg Tiedemann. Parallel corpora in Linköping, Uppsala and Göteborg (PLUG): The corpus. Working Papers in Computational Linguistics and Language Engineering 14, Department of Linguistics, Uppsala University, Sweden, 1999.

[Tie99c] Jörg Tiedemann. Word alignment - step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics (NODALIDA)*, pages 216–227, University of Trondheim, Norway, 1999.

[Tie01a] Jörg Tiedemann. Can bilingual word alignment improve monolingual phrasal term extraction? *Terminology International Journal of Theoretical and Applied Issues in Specialized Communication*, 7(2):199–215, 2001.

[Tie01b] Jörg Tiedemann. Predicting translations in context. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 240–244, Tzigov Chark, Bulgaria, 5-7 September 2001.

[Tie01c] Jörg Tiedemann. UplugWeb - corpus tools on the web. presented at the 13th Nordic Conference on Computational Linguitics, (NODALIDA), Uppsala, Sweden, 2001.

[Tie01d] Jörg Tiedemann. The use of parallel corpora in monolingual lexicography - How word alignment can identify morphological and semantic relations. In *Proceedings of the 6th Conference on Computational Lexicography and Corpus Research (COMPLEX)*, pages 143–151, Birmingham, UK, 28 June - 1 July 2001.

[Tie02a] Jörg Tiedemann. MatsLex - a multilingual lexical database for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, volume VI, pages 1909–1912, Las Palmas de Gran Canaria, Spain, 2002.

[Tie02b] Jörg Tiedemann. Uplug - a modular corpus tool for parallel corpora. In Lars Borin, editor, *Parallel Corpora, Parallel Worlds*, pages 181–197. Rodopi, Amsterdam, New York, 2002. Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden, 1999.

[Tie03] Jörg Tiedemann. Combining clues for word alignment. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 339–346, Budapest, Hungary, April 2003.

[TIM02] Kristina Toutanova, H. Tolga Ilhan, and Christopher D. Manning. Extensions to HMM-based statistical word alignment models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pages 87–94, 2002.

[TKS96] Erik F. Tjong Kim Sang. Aligning the Scania corpus. Technical report, Department of Linguistics, University of Uppsala, 1996.

[TKS02] Erik F. Tjong Kim Sang. Memory-based shallow parsing. *Journal of Machine Learning Research*, 2:559–594, March 2002.

[TN03] Jörg Tiedemann and Lars Nygaard. Opus - an open source parallel corpus. In *Proceedings of the 13th Nordic Conference on Computational Linguistics*, University of Iceland, Reykjavik, 2003. to appear.

[TRA03] TRADOS. *WinAlign User Guide*. Dublin, Ireland, 2003.

[vdE93] Pim van der Eijk. Automating the acquisition of bilingual terminology. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 113–119, Utrecht/The Netherlands, April 1993.

[Vér98] Jean Véronis. *ARCADE: Tagging guidelines for word alignment*. Laboratoire Parole et Langage, Université de Provence & CNRS, Aix-en-Provence Cedex 1, France, April 1998.

[VL00] Jean Véronis and Philippe Langlais. Evaluation of parallel text alignment systems. the ARCADE project. In Jean Véronis, editor, *Parallel Text Processing*, Text, speech and language technology series, chapter 19. Kluwer Academic Publishers, Dordrecht, August 2000.

[VNT96] Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Confernece on Computational Linguistics (COLING '96)*, pages 836–841, Copenhagen, Denmark, 1996.

[VOT+00] Stephan Vogel, Franz Josef Och, Christoph Tillmann, Sonja Nießen, Hassan Sawaf, and Hermann Ney. Statistical methods for machine translation. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 377–393. Springer Verlag: Berlin, Heidelberg, New York, Berlin, July 2000.

[Wu94] Dekai Wu. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 80–87, New Mexico State University, 1994.

[WX94] Dekai Wu and Xuanyuin Xia. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA)*, Columbia, MD, 1994.

[YNW01] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st International Conference on Human Language Technology Research (HLT)*, 2001. available from http://citeseer.nj.nec.com/yarowsky00inducing.html.